Active Learning for drug discovery

Kevin Williams

Institute of Mathematics, Physics and Computer Science

Aberystwyth University

11 March 2014

This thesis is submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy of Aberystwyth University

Declaration

This thesis has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed:

Date:

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed:

Date:

Statement 2

I hereby give consent for my thesis, if accepted, to be made available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed:

Date:

Statement 3

I hereby give consent for my thesis, if accepted, to be deposited in the University's Institutional Research Repository.

Signed:

Date:

Abstract

This thesis describes work conducted to enable Robot Scientist Eve to autonomously evaluate drug-like chemicals during high throughput experiments. Eve tests libraries of chemical compounds against yeast-based targets expressing parasite and host (human) proteins (i.e. DHFR, NMT & PGK); the parasites included in this study are responsible for an array of neglected tropical diseases.

The raw data for yeast growth curves from an initial screen were evaluated, and decision tree rules were constructed to describe the relative activity and toxicity of compounds. These rules were verified, and versions were subsequently developed for application to routine mass and confirmation screens. Consequently, many potential lead drug-like candidates have been identified in the Maybridge Hitfinder library; several compounds from an approved drug library (the Johns Hopkins Clinical Compound Library) have also been confirmed as exhibiting activity against these yeast-based targets. Further *in vivo* study of some JHCCL compounds is in progress using extracted parasite proteins; preliminary results indicate the potential for repositioning Triclosan and Tnp-470 as having anti-malarial behaviour based on their interaction with *Plasmodium sp.* DHFR proteins.

In the second phase of the programme, a prototype Active Learning strategy was applied (*active k-optimisation*) to partial mass screen data as a seed; this allowed Eve to select compounds by assessing and predicting quantitative structure activity relationships (QSAR) between seed and unknown compounds. Simulations of learning and testing QSAR cycles showed that Eve would be able to select active compounds more efficiently under such a regime. Other strategies have been developed that further improve selection efficiency for active compounds, and also promote the ability to find rare category compounds. An econometric model has been developed to demonstrate the potential beneficial impact of Active Learning strategies on the execution costs for such screens.

Acknowledgements

I would first like to thank my friends from before Aberystwyth for their support and encouragement in taking this route, especially Meryl and Nathalie for their insight, and the late Adrian Dunn for his kind words.

I would like to thank my primary supervisors: Professor Ross King for his continued support and motivation throughout my studies, and Dr Chuan Lu for stepping into Ross' official role when he moved on to Manchester University. I would also like to thank the other members of the Robot Scientist team, especially Andrew Sparkes, Jem Rowland and Ron Pateman for their assistance and ideas, and Bessie Bilsland and Kurt De Grave for raising good questions.

I have met many interesting people since coming to Aberystwyth, and made many friends who have helped me through the last four years. I would also like to thank the friends I've met on my bi-annual trips to Lundy, a useful bolt-hole to escape the pressures of everyday life. Each of these relationships has helped me in some way to negotiate the tricky, rocky patches lurking just out of site from my terminal.

Finally, I would like to acknowledge the partial funding provided by Pfizer, and the waiving of my second and third year fees by the Computer Science department.

Contents

List of Tables

List of Figures

List of Equations

Prologue

"*Epilogue:*

One of the greatest benefactors of all lifekind was a man who couldn't keep his mind on the job in hand.

Brilliant?

Certainly.

One of the foremost genetic engineers of his or any other generation, including a number he had designed himself?

Without a doubt.

The problem was that he was far too interested in things which he shouldn't be interested in, at least, as people would tell him, not now.

He was also, partly because of this, of a rather irritable disposition.

So when his world was threatened by terrible invaders from a distant star, who were still a fair way off but travelling fast, he, Blart Versenwald III (his name was Blart Versenwald III, which is not strictly relevant, but quite interesting because --- never mind, that was his name and we can talk about why it's interesting later), was sent into guarded seclusion by the masters of his race with instructions to design a breed of fanatical superwarriors to resist and vanquish the feared invaders, do it quickly and, they told him, ``Concentrate!''

So he sat by a window and looked out at a summer lawn and designed and designed and designed, but inevitably got a little distracted by things, and by the time the invaders were practically in orbit round them, had come up with a remarkable new breed of super-fly that could, unaided, figure out how to fly through the open half of a half-open window, and also an off-switch for children. Celebrations of these remarkable achievements seemed doomed to be shortlived because disaster was imminent as the alien ships were landing. But astoundingly, the fearsome invaders who, like most warlike races were only on the rampage because they couldn't cope with things at home, were stunned by Versenwald's extraordinary breakthroughs, joined in the celebrations and were instantly prevailed upon to sign a wide-ranging series of trading agreements and set up a programme of cultural exchanges. And, in an astonishing reversal of normal practice in the conduct of such matters, everybody concerned lived happily ever after.

There was a point to this story, but it has temporarily escaped the chronicler's mind. "

From 'So long, and thanks for all the fish', Douglas Adams, 1984.

Chapter 1

Introduction

Adventures with yeast, part 1 of 7: Khorasan loaves

Dough starter:

5 grams	*fresh yeast*
135 ml	*cold water*
100 grams	*strong white bread flour*
100 grams	*Khorasan flour*

Whisk the yeast into the water. Add the flours and mix, then store in a warm place for at least six hours.

Main dough

680 ml	*warm water*
940 grams	*strong white bread flour*
130 grams	*Khorasan flour*
15 grams	*sea salt*
22 grams	*fresh yeast*

Mix the water into the starter. Add this mixture to the flours and other ingredients, and knead for 10 minutes. Allow it to rise in a warm place for one hour or longer, then knock back, shape into two loaves, and leave to rise again for 30 minutes. Dust with flour, and slash the top of the loaves with a sharp knife. Bake in a preheated oven at 230°C for 40 to 45 minutes.

1.1 Background

The process for drug discovery and development is expensive and of high risk, and largely the preserve of pharmaceutical companies (**Morgan *et al*., 2011**). Whilst independent and specialised discovery research groups do exist in small companies and academia, the long and expensive registration process almost invariably means that larger companies usually step in once potentially successful candidate drugs have been identified.

The underlying problem in the discovery phase is to find lead compounds that maximise the probability of discovering an effective pharmaceutical. This problem is not only limited by scientific knowledge and its application, but is also strongly affected by resource management, balancing the costs of time and money associated with this process against the expected benefits of profit and health improvements.

Simple candidate selection of drug-like compounds is linear and iterative, where compounds from a large library are individually tested against a target based on a micro-organism to find those with notable activity. An alternative approach is to make use of quantitative structure-activity relationship (QSAR) techniques where existing seed/lead chemical compounds are used as indicators of possible alternative candidates. Once lead compounds have been identified their structures can be scrutinised by an expert, and modified if necessary to further improve their drug-like properties.

It is proposed that a further enhancement would be to have a continuous learning process to incorporate all information obtained during the experiments, thereby enabling re-modelling of the selection process as more data becomes available. This becomes a problem suited to Active Learning (AL) techniques, where AL is a branch of machine learning in which algorithms are designed to continuously select the next best examples to test. AL has previously been applied only sparingly to QSAR analysis (**Warmuth *et al*., 2003; De Grave *et al*., 2008a**), most likely due to insufficient public domain compound/target data available for effective models to be built. The proposed development programme for Robot Scientist Eve **(Sparkes *et al*., 2010)** explicitly included AL methods as part of its structure, with one aim being that Eve would conduct fully automated drug screening and discovery experiments.

1.2 The Robot Scientist project

Originally based in Aberystwyth University, with Professor Ross King as the founder and Principal Investigator (**King *et al.*, 2009**), the Robot Scientist project moved to the Manchester Interdisciplinary Biocentre (MIB), University of Manchester in 2012.

The concept is to fully automate scientific method by linking up and automating the steps that form the scientific discovery process, applying rules and procedures with absolute objectivity and without bias; a Robot Scientist should be able to generate hypotheses, devise and run repeatable experiments to test them, interpret the results and refine each hypothesis. This cycle is repeated until a satisfactory outcome is reached, and all methods and results are recorded in order that the work can be reproduced by others.

Adam was the first Robot Scientist in the programme; it was a combination of laboratory worker and data analyst, and has shown its capacity for independent hypothesis-led experimentation by discovering novel gene function in yeast strains (**King *et al.*, 2004**). The system was designed to run microbial growth experiments to explore the amino acid pathway of the yeast *Saccharomyces cerevisiae*, with all liquid handling, growth measurements and analyses as linked up automated steps, requiring minimal human intervention after the initial set up.

Robot Scientist Eve has been developed specifically to run high throughput experimentation (HTE) for testing the activity of libraries of drug-like compounds against genetically modified yeast assays (**Sparkes *et al.*, 2010**); the yeast strains used to date have been developed by the School of Biological Sciences, University of Cambridge **(Bilsland *et al.*, 2011)**. The strains have been modified to include enzymes from parasites responsible for certain neglected tropical diseases; these targets are evaluated alongside orthologue human strains, with the aim of identifying differential activities and possible toxicity.

1.3 Drug-like compound selection processes

Mass screening work conducted on Eve is limited to using a collection of drug-like compounds provided by the Maybridge Hitfinder library (www.maybridge.com); this collection is small compared to commercial libraries but is designed to be chemically diverse, covering a wide range of pharmaceutical functionality. Analysis of the large array of chemical and biological assay data generated in each screen has needed to be formalised as part of this thesis, and has been fundamental in developing Robot Scientist Eve as a prototype drug discovery system. These data are then used to identify potential drug-like activity.

When isolating lead drug-like compounds it would be ideal if the next selected candidate showed strong activity with minimal toxicity to the human host, but in early stages strong candidates with notable toxicity might be also of value as indicators of similar structures with potentially strong activity. However, when building and applying Active Learning routines, there are several other criteria to consider when selecting the next untested candidate compound within the screening process: e.g. should it be the one with maximum predicted variance, the maximally optimistic one, the one with maximum predicted mean value, the one offering the maximum predicted improvement? All these criteria have conflicting effects, and building a balanced AL model depends on consideration of their relative importance.

It was planned that Eve's prototype AL selections would use the *active k-optimisation* strategy (**De Grave *et al*., 2008a**), and be compared to Pfizer's Naive Bayesian approach (company confidential). These could be benchmarks for additional models, to be designed using simple chemical knowledge. Finalised models should select the next n best candidates from a list of untested compounds, whilst allowing the model to develop and search the wider chemical space. The ability to search much larger libraries must also be considered, together with an eye on exploring the space outside constraints applied by the library; this latter aspect might further develop the Robot Scientist project, where activity predictions for novel molecules could be made before they are synthesised and tested.

1.4 Challenges and goals

The primary intent for this thesis was to find novel ways to carry out drug discovery; the scope of this work was beyond simply providing support to Eve's development, and it was intended that AL strategies devised would have much more widespread applicability. The final strategies were based on simple inputs (standard chemical fingerprint techniques, and classification of drug-like activity) that could readily be applicable to many systems or problems. Robot Scientist Eve was to provide the data sets upon which these methods would be developed.

The novelty of Eve and the associated yeast strain targets meant that very specific problems would need to be solved before moving on to develop the Active Learning algorithms that would satisfy the primary intent.

The verification of several aspects of Eve's processes eventually fell under the remit of this thesis:

- It needed to be shown that the growth of the yeast-based targets could be measured consistently and objectively in a high throughput system.
- Discrimination needed to be proven between active and inactive treatments, as provided in the form of positive and negative controls.
- The results generated by Eve needed to be shown to be consistent and significant on a continuous basis across all screens; this would show that Eve is capable of consistently running good quality high throughput experiments.

Once good quality datasets had been built, the Active Learning phase would hold the following challenges:

- The prototype methods for Active Learning needed to be shown as offering improvements over simple, linear, methods for candidate identification, and a means of measuring these benefits needed to be designed.
- Alternative AL approaches needed to be found that might offer further enhancements over the prototype method.

These AL methods would be tested using Eve's data, but would be designed with general results in mind to enable wider applications.

In addition, the combined effect of meeting the above targets would allow Eve to operate as a standalone entity for the drug discovery process, whilst also compiling specific empirical datasets for the drug-like compounds/target combinations that might prove useful to other researchers.

1.5 Thesis organisation and structure

This thesis follows the pathways of the processes built for Eve, initially for isolating and identifying activity levels against different modified yeast targets, then for the tools that enable the Active Learning routines to be designed and evaluated.

Chapter 2 is a review of the building blocks required in conventional drug discovery processes. The general nature of pharmaceutical research is discussed, together with current techniques such as QSAR (quantitative structural activity relationships). Details are given of the yeast targets, and the parasites responsible for the neglected tropical diseases for which they will provide analogues. This chapter includes an examination of the types of tool available to the computational chemist, albeit from a perspective based on open source software, and will start to identify areas where Robot Scientist Eve group might provide benefits. Machine learning and other computational techniques are reviewed, with a bias towards modelling datasets from chemical problems

Active Learning techniques are described, together with their possible application to drug discovery. Variations that can boost the performance of AL are investigated, with a leaning towards finding regions of the chemical space less oft explored.

In Chapter 3 the raw data generated by Eve is analysed, and rules are built to extract and simplify the relative activity of compounds. Mass screen data are used to generate candidates for confirmatory testing, and lists of active candidates are built. Work conducted on existing drug therapies from the Johns Hopkins Clinical Compounds Library is reported

Chapter 4 devises a scheme for using mass screen data as a proxy for confirmation screen data, thereby allowing *in silico* simulations of Active Learning algorithms. AL processes are refined to incorporate either numerical or classification data, thereby

expanding the options away from the prototype AL method. Commentary on different approaches is provided, with variations in clustering methods, transfer learning and rare category detection being considered.

Chapter 5 reports the results of the simulations for each AL method, and makes a detailed comparison of their performances.

Chapter 6 expands on the AL simulation studies by incorporating an econometric model. The potential for improving the efficiency of drug discovery is discussed when using data handling processes similar to those developed using Eve's experiments.

Chapter 7 summarises the highlights of the work conducted in support of this thesis. Conclusions are drawn on the approaches used, and ideas are provided for further work and possible additional publications.

Appendix A provides data drawn from mass and confirmation screens reported in Chapter 3; it includes lists of all compounds with suggested activity versus the yeast analogues of parasitic organism enzymes.

Appendix B is a compilation of all simulations run using the Active Learning algorithms described in Chapter 5.

Appendix C gives details of the data file structure provided by Eve, and the hardware and software conformations used for the data analyses and simulations.

1.6 Contributions to knowledge, and related work

Contributions to knowledge

Details of the main achievements of this programme are given in Sections 7.1 & 7.2.

Active Learning strategies based on classification of drug-like activity have been developed successfully; by using simple input data, these strategies should be readily adaptable to other drug discovery regimes, and will also have applicability for dealing with other problems where rare category detection is required.

Specifically, this work has enabled Robot Scientist Eve to autonomously evaluate drug-like chemicals, moving between mass screen and confirmation screen modes.

It has also been shown by simulation that Eve might switch from mass screening to an intelligent screening mode, with distinct improvement in the rate of detection of active compounds. Similarly, it has been shown that active compounds with rare structures (i.e. dissimilar to compounds known to be active) can be promoted and found at earlier stages compared to simple search protocols.

Eve's confirmation work on existing drug therapies suggests potential activity against neglected parasitic diseases for 14 of these, and also suggests a possible mode of parasite growth inhibition for two candidates with previously identified *in vivo* effects.

Publications

The work on the Robot Scientist Eve programme was ring-fenced in order to support major publications, hence it was not possible to publish intermediate studies.

A poster (The Robot Scientist Eve: selection and confirmation of chemical compounds with potential for activity against parasites causing neglected tropical diseases, Williams *et al*.) was presented at the SLAS exhibition in San Diego in February 2012. This was supported by an academic travel award from SLAS under their Tony B. Award programme for new researchers.

Some results from the data analysis for Robot Eve have now been published in (**Bilsland *et al*., 2013**).

The work on applied machine learning for mass & confirmation screens, and the prototype *active k-optimisation* algorithm is in an advanced draft; this publication will include evidence of possible drug repositioning for JHCCL compounds.

A further publication with specific drug repositioning for Triclosan is also in an advanced draft.

The two papers currently under revision were originally submitted in the form of a single publication to *Science*:

Cheaper Faster Drug Development Validated by the Targeted Repositioning of Triclosan against Dihydrofolate Reductase in Malarial Parasites (Kevin Williams, Elizabeth Bilsland, Andrew Sparkes, Wayne Aubrey, Michael Young, Larisa N. Soldatova, Kurt De Grave, Jan Ramon, Liisa Van Vliet, Jack E. Feltham, Florian Hollfelder, Michaela de Clare, Worachart Sirawaraporn, Victoria Jackson, Stephen G. Oliver, Ross D. King).

Chapter 2

Background knowledge for drug discovery and Active Learning

Adventures with yeast, part 2 of 7: Ninja wine

1 kg	*fresh root ginger*
5 kg	*white sugar*
1 kg	*sultanas*
3	*juiced lemons*
1 cup	*strong black tea*
1 teaspoon	*Marmite*
1 packet	*desert wine yeast*

Whilst generally following all the normal steps for home wine making...

Macerate or finely chop the ginger, and add to 5 litres of water. Heat the mixture to near boiling point, then allow it to cool to room temperature over a couple of hours.

Add the marmite and 3 kg of sugar to 5 litres of water; heat this until the sugar is dissolved, then allow to cool to room temperature.

When cooled, combine the two mixtures in a sanitised container. Add the lemon juice and tea. Macerate the sultanas and add these to the mixture with stirring. Make up to 15 litres with cold water. Sprinkle the yeast onto the mixture, and mix in after 15 minutes.

After two days of fermentation, add 1 kg of sugar to the mixture with stirring. After two further days, add the remaining sugar with stirring.

2.1 Drug development from academia's perspective

Success in the field of drug design can be readily quantified, but less readily predicted. The drug development process is exposed to rigorous approval steps, and failure is common at each way point; a pharmaceutical company may start trials with several tens or hundreds of compounds to treat a certain condition, with the hope of finding one commercially feasible molecule.

A candidate compound will need to pass through a minimum of the following steps:

- Finding the chemical lead compounds (*in silico, in vitro*)
- Pre-clinical trials: testing for absorption, distribution, metabolism, and excretion (ADME) properties (*in vitro, in vivo*)
- Clinical trials: Phase I (safety), II (effectiveness) & III (definitive) trials (*in vivo*)
- There may also be Phase 0 (micro dose in humans to understand pharmacokinetic properties) and Phase IV trials (post-approval monitoring).

The full process will take many years and many $100m to complete for a single therapeutic agent. Selection of the starting point of the work will have a strong influence on the relative success of the development programme.

The power of computational techniques has improved over time, resulting in a shift in the methods for finding lead compounds from *in vitro* experimentation towards *in silico* work. **(Goodford, 1984)** gave a synopsis of the previous two decades' work in drug design using a receptor fit approach, and also pointed to potential future developments (e.g. drugs tuned to the individual requirements of patients based on their genetic code).

Christopher Lipinski's (of Pfizer) work has been pivotal in identifying fundamental properties expected of drug-like compounds, and mooted a series of rules for drug discovery when using computational and experimental modelling **(Lipinski *et al*., 1997)**. He argues that the empirical approach to drug design has been replaced by rational approaches, largely due to extensions in the knowledge base from HTE. *In vitro* HTE has allowed a dramatic change to the previous process of lead compound selection consistent with historically orally active compounds; large libraries can now

be examined initially for hits, with screening for other properties such as IC_{50} (inhibition concentration of a drug at which biological activity is halved) and solubility now taking place at a later time. Subsequently, new lead compounds do not necessarily fit the classical "active" profile *in vivo*. Other properties now come to the fore, allowing a wider scope of compounds to be tested *in vitro* thanks to improved solvation systems **(Patel and Gordon, 1996)**. Any leads from this work can be examined for relevance and possible modification to improve their usability. One downside of this approach is that very active, but otherwise unsuitable, compounds might eclipse compounds with lower potency but favourable therapeutic profiles.

2.2 Organic chemical structures and general drug-like properties

The predictive models for drug development (see sections 2.10 & 2.11 for an expansion of these techniques) make use of a wide range of structural and physicochemical parameters to describe the behaviour of seed and candidate molecules. Earlier predictive work used simpler datasets, largely using 2D structural modelling techniques in addition to activity and solubility measurements such as:

IC_{50}	Half maximal inhibitory concentration: the concentration of a drug at which the activity of a biological process is halved. High potency is much preferred, as this minimises the impact of weaker factors elsewhere.
Molecular weight	Small molecules are preferred at the early stages of development, as they have a tendency to diffuse more effectively after application.
Solubility	Drug-like molecules need to be transported in the aqueous blood phase, and be able to pass through lipid cell membranes.
Lipophilicity	A measurement of relative solubility in the lipid phase, using the logarithm of the octanol:aqueous solubility partition coefficient ($log\ P$).
$C\ log\ P$	A computationally predicted version of $log\ P$.
$M\ log\ P$	A measured version of $log\ P$ from experiments.

H-bond donor — Increased numbers of hydrogen bond donors in a molecule generally improves aqueous solubility at the expense of lipid solubility, and helps to define the relative ease with which the compound passes through different membranes, e.g. cell walls; blood-brain barrier.

More recently, it is computationally feasible that several hundred parameters might be measurable or predicted for each molecule and its activity within an experiment, and these data used in the model. Complex 3D structures of candidate molecules and targets may now be constructed, in order to predict interactions.

2.3 Discovery methods based on drug-like properties

Compounds listed in the Derwent World Drugs Index (WDI) that reached Phase II trials have been systematically analysed for parameters influencing efficacy: molecular weight, lipophilicity (octanol solubility:aqueous solubility) as $C\,log\,P$ and/or $M\,log\,P$, hydrogen bond donor and acceptor groups. Analysis of these parameters led to generation of "The Rule of 5", such that poor absorption and permeation are more likely when:

- There are more than 5 hydrogen bond donors (sum of –NH and –OH)
- The molecular weight is greater than 500
- $log\,P$ > 5 (or $M\,log\,P$ > 4.15)
- There are more than 10 hydrogen bond acceptors (sum of Ns and Os)

Some compound classes contain notable exceptions to "The Rule of 5", e.g. substrates for biological transporters (antibiotics, antifungals, vitamins, cardiac glycosides).

Compounds reaching Phase II trials that had exceeded two or more parameters occupied at most 10% of the data set for any combination. If a drug-like compound doesn't fit these rules it is likely to need a high activity to be useful (a function of dose, solubility and permeability to describe potency); the Rule of 5 doesn't allow for this factor as it is based on simplified limits rather than such combination effects.

The "Rule of 5" was extended to describe the properties required of an effective pharmaceutical molecule and the distribution of families of such compounds across

the whole domain of medium sized organic chemicals **(Lipinski, 2000)**. At that time, the number of drug-like compounds in existence was estimated at 10000, with an estimated number of targets as 500 strong **(Drews, 2000).**

In theory there are more than 10^{50} molecules up to molecular weight 600 that contain the atoms commonly found in drugs **(Lipinski, 2000)**; this is an immense space, but it is Lipinski's opinion that pharmacologically active compounds only occupy discrete areas in this space, and that random screening across it amounts to a lottery. Drug companies duly take a conservative approach, and stay close to existing knowledge families; this suggests that alternative rational searches of other areas of the chemical space might be beneficial if suitable starting points could be identified.

(Lipinski, 2006) also discusses a demarcation between how commercial and academic research organisations should go about their work. He argues that, almost by definition, pharmaceutical and biotechnology companies are unlikely to take large strides into unknown territory as there is too much risk for it to be fruitful; this leaves such areas open for academia to investigate, and yet there is seemingly insufficient support for them to do so. The cycle by which academia acquires research funding and generates papers seems to produce an almost equally conservative approach; the push for universities to claim chunks of intellectual property is seemingly compounding this effect, with the academic now having further bureaucratic hurdles to jump, and making external relationships more difficult to forge.

Lipinski suggests that the main areas of academic interest for drug screening/design should be those away from the mainstream, or in areas where there is little or no potential profit (e.g. Neglected Tropical Diseases); he argues that university Intellectual Property officers may not realise that trying to protect IP in neglected, low profit areas is fairly pointless, and its pursuit will only harm potential partnerships.

An area that should be of most interest to academia is the use of small molecules for verification of screening tests and/or investigating biological pathways, and these tasks are far removed from drug discovery. It is in this type of work that "big pharma" and academia need to have better links as the former have large databases of what doesn't work (and why) and can help to eliminate studies doomed to later difficulties; this needs openness and trust in both directions, which in turn is an approach that is becoming increasingly difficult to foster.

2.4 Parasites responsible for 'Neglected Tropical Diseases'

Neglected Tropical Diseases (NTD) are largely defined and described by chronic impacts in impoverished communities through their debilitating effects on health and development (**Feasey *et al.*, 2010**). The term is effectively a catch-all for diseases that have lain outside the scope of development programmes of pharmaceutical companies, largely due to limited impacts on targeted commercial markets and hence limited immediate profitability.

The last decade has seen greater recognition of the need to combat NTDs, and to assist those communities affected by helping with prevention, education, control and treatment. Traditional academic funding in this area is increasingly augmented by national and philanthropic bodies, e.g. U.S. National Institutes of Health (NIH); Bill & Melinda Gates Foundation; Wellcome Trust (**Moran, 2011**), and dedicated high profile journals also now exist (e.g. PLOS Neglected Tropical Diseases, established in 2007).

The infectious parasitic diseases within the scope of the Robot Scientist programme are caused by either protozoa (unicellular eukaryotic organisms) or helminths (parasitic worms); existing therapies suffer from restrictions due to varying combinations of increased drug resistance, low effectiveness, difficult side effects, and high expense. Current information for each of the parasites/diseases can be gleaned from (www.who.org).

Malaria (wild and drug-resistant strains of *Plasmodium falciparum* and *P.vivax* in this study) is a major problem for developing countries with an estimated 219 million cases in 2010, leading to ~660,000 deaths, mostly among African children. Existing drug therapies are effective for the wild-type strains, but increasing incidence of drug resistance (artemisinin and derivatives) is a growing concern (**Dondorp, 2009**).

The protozoa responsible for Chagas disease, *Typanosoma cruzi* (**de Souza *et al*., 2010**), is endemic in a large reservoir of wild animals in Central and South America, and the parasite cannot be eradicated. Infection is estimated at 7 to 8 million people, mostly in Latin America. Whilst effective treatments exist, they need to be employed soon after infection and have adverse side effects in a significant number of patients (reported up to 40%).

Sleeping sickness (Human African trypanosomiasis) is transmitted by the bite of infected tsetse flies in sub-Saharan Africa. Two subspecies of the *Trypanosoma brucei* parasite cause the disease (**Steverding, 2010**): untreated, those infected with the far commoner *T. b. Gambiense* have a possible survival time of several years, whereas *T. b. Rhodesiense* will cause death within months. Different treatments are required at the first (asymptomatic) and second stages, with those effective in the latter being significantly more toxic.

Leishmaniasis is transmitted by bites from the infected female phlebotomine sandfly. There are some twenty species of *leishmania* protozoa, causing various problems ranging from the fatal visceral form (caused by *L. Donovani*) to skin lesions (**Melby *et al.*, 1992**). 1.3 million cases are estimated *per annum*, with 20-30,000 deaths.

Schistosomiasis is caused by *Schistosoma spp*. blood flukes. In sub-Saharan Africa, more than 200,000 deaths are due to infection with *S.mansoni*. Praziquantel is an effective treatment for all forms of Schistosomiasis, but is seemingly only available to 12% of the estimated 243 million infected people (2011 figures), and there are concerns that drug-resistant parasites may develop (**Melman *et al.*, 2009**).

2.5 Protein targets and yeast strains

For this project, the yeast *Saccharomyces cerevisiae* has been genetically engineered to contain drugable enzyme targets from the parasite strains. The yeast strains for Eve were developed by Dr E Bilsland, School of Biological Sciences, University of Cambridge (**Bilsland *et al.*, 2011**).

These modified yeasts have been used as targets to counter significant problems that would arise in high throughput drug screening if whole parasites were otherwise under test. Three drugable protein targets have been used for this work:

- Dihydrofolate reductase (DHFR) is the enzyme responsible for conversion of dihydrofolate into tetrahydrofolate (**Bertino, 2009**), an essential process in the growth of all organism types.
- Phosphoglycerate kinase (PGK) is an essential enzyme for parasites in their development stage in the host's blood (**Michels *et al.*, 2006**).
- N-myristoyltransferase (NMT) is a protein modifier, which allows parasites to target certain membranes (**Frearson *et al.*, 2010**).

The human and parasite protein strain have been used to replace the equivalent function in wild-type yeast.

The three strains chosen for each assay needed to fluoresce at discrete wavelengths in order to be measured without interference; this was achieved by Dr Bilsland by attaching different fluorophores to the strains. The three fluorophores used to date are labelled as mcherry (580 nm excitation wavelength, 612 nm emission), sapphire (405/510 nm) and venus (500/540 nm). Initial trials with Eve used pairs of yeast strains, but quickly progressed to the triple strain approach used for the main body of the experiments. It was proposed that Eve might operate using four strains with discrete fluorophores, but this has not been attempted to date.

Gene	**Organism strains**
DHFRpdr5	Hs - Human Pv - Plasmodium vivax PvR - Drug resistant P.vivax Pf - P.falciparum PfR - Drug resistant P.falciparum Sm - Schistosomiasis mansoni Tb - Trypanosoma brucei Tc - Trypanosoma cruzi Lm - Leishmania major
PGKpdr5	Hs, Pv, Sm, Tb, Tc
NMTpdr5	Hs, Pv, Sm, Tb, Tc

Table 2.1: Modified target yeast strains for Robot Eve

The drug resistant PvRdhfr strain is a triple mutant for residues S58R, S117N, and I173L; the resistant PfRdhfr strain is a triple mutant for N51I, C59R & S108N. Several other drug resistant double, triple and quadruple mutants of PvDHFR and PfDHFR are known in the field **(Cowman *et al.*, 1988; Sirawaraporn *et al.*, 1997).**

2.6 Mutations and drug resistance

Drug resistance in parasites occurs through selection in the population. The drug removes those strains susceptible to its actions, leaving the field clear for less susceptible strains of the organism to flourish. If the drug target site embedded in the structure of the protein is obscured by mutation, the ligands from the drug might be blocked from binding. Within any population, parasites will exist where these mutated structures occur, but under normal circumstances will be out-competed by the more successful, unmutated organism (by definition). As an example, a study of relative effectiveness of antifolates in a wide range of examples of *P.vivax* where the DHFR protein is mutated has shown that enzyme efficiency is distinctly adversely affected with increased mutation (**Auliff *et al*., 2010**); these mutated strains would easily be out-competed by unmodified examples.

Drug resistance in *P.falciparum* is a serious problem. The anti-folate mechanism by which pyrimethamine is an effective drug, was shown to be disrupted by mutation of the DHFR structure (**Peterson *et al*., 1988**). The range of artemisinin-based antimalarials is also reported to be experiencing resistance in some areas of South East Asia (**Dondorp *et al*., 2009**).

For *P.falciparum*, the structure of the DHFR domain consists of ~198 amino acids, and the wild-type drug-sensitive strain is labelled as 3D7. A mutant version having pyrimethamine resistance (labelled HB3) exists where a single amino acid residue at position 108 has changed from serine (S) to aspargine (N), and is thus referred to as the S108N mutant.

Other prevalent double and triple mutant strains are **(Cowman *et al*., 1988)**:

7G8 double mutant	(S108N, N51I)	
*Csl-2 triple	(S108N, C59R, I164L)	Thailand
*K-1 double	(S108N, C59R)	Thailand
*V-1 double	(S108N, C59R)	Vietnam
Palo-Alto double	(S108T, A16V)	Uganda

The strains marked * were ~100 times more resistant to pyrimethamine than 3D7, with the others showing intermediate resistance. The Palo-Alto strain is resistant to cycloguanil (**Sirawaraporn *et al*., 1997**), relating to the local use of this drug.

Work on alternative therapies has focussed on identifying the changes to the DHFR structure, and proposing alternative ligands. One earlier interesting compound, WR99210, was found to have activity in various mutations of the PfDHFR protein, including a quadruple mutant **(Yuvaniyama *et al*., 2003**). Another compound, QN254, also showed promise against such targets, but later mouse and rat modelling suggested too small a therapeutic window **(Nzila *et al*., 2010)**.

Similarly, work on *P.vivax* strains has indicated that the para-chlorine atom on pyrimethamine causes steric hindrance in the double mutant SP21 (S58R, S117N; note: this corresponds to C59R and S108N in *P.falciparum*). An analogue of pyrimethamine with the chlorine atom removed has a seven-fold improvement in the inhibition of SP21 **(Kongsaeree *et al*., 2005).** Further knowledge of modes of mutation might assist with the idea of using multiple therapies with opposing modes of selection (**Ridley, 2002**).

For the experiments on Eve, the drug resistant PvR DHFR strain is a triple mutant (S58R, S117N, I173L); the resistant PfR DHFR strain is a triple mutant (N51I, C59R, S108N) (**Bilsland *et al*., 2013**).

Studies of *leishmania* and *trypanosoma spp*. (**Chakravarty and Sundar, 2010; Baker *et al.*, 2013**) have identified drug resistant mutations, and other work on helminths (**James *et al.*, 2009**) suggests that both changes in genes and their expression are occurring, allowing the organism to adapt, evolve and survive.

2.7 2D representations of organic chemical structures

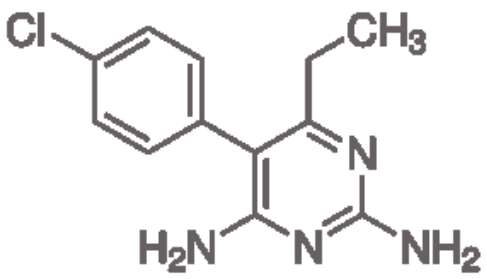
The methods used in this project for representing chemical structures (SMILES codes), and the tools used for similarity searching across the available chemical space (Tanimoto Similarity coefficients) were chosen to fit in with those used for the prototype Active Learning method. This approach was decided to reduce the level of variability when comparing different AL algorithms, as it these which were to be examined rather than the overall efficacy of the compound selection processes.

The scaffold structure of an organic chemical lends itself to be codified. The presentation of atoms and bonds follow physical rules; the forces and interactions within a molecule can be calculated, which in turn allow 3D representations to be built. However, in terms of the predictive modelling requirements for HTE, a 2D representation of molecular structure is far simpler to work with; SMILES codes, SMARTS patterns (**James *et al*., 1997**), InChI strings (**Heller and McNaught, 2009**), MACCS keys and ILP processes are amongst variants which have successfully helped to model chemical compound activity within groups of molecules.

SMILES codes (Simplified Molecular Line Entry System)

The 2D representation of an organic molecule can be codified into a linear sequence that represents the atoms, bonds and chirality of its structure. SMILES are one such means of depicting this information. The initial work on producing the SMILES specification for coding organic structures was conducted in the 1980s (**Weininger, 1988; Weininger *et al*., 1989**). This work was built on extensively and was launched as an open standard in 2007, along with a large resource of web-based software for chemical informatics (**Tetko et al, 2005**; **Tetko, 2005**; **Guha *et al*., 2006**). There is a significant amount of prior experience in chemoinformatic analyses that show the usefulness of using SMILES codes as a generalised coding system, and its adoption has been fairly widespread. SMILES also offer benefits over key-based systems developed for pharmaceutical studies (e.g. MACCS, InChI) due to their adaptability; it can be speculated that this might allow identification of unusual patterns that might not be seen when using highly specialised tools, and raises the case for a combined approach.

Another open source project, Open Babel (**Babel**), allows the chemist to search, convert, analyze, or store data from molecular modelling, chemistry, solid-state materials, biochemistry, or related areas. One application available in Open Babel is to convert SMILES codes into a fingerprint (Daylight FP2), generating an output vector that describes fragments that make up the molecule.



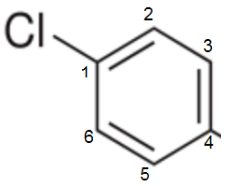
Molecular formula: $C_{12}H_{13}ClN_4$

SMILES code: Clc2ccc(c1c(nc(nc1CC)N)N)cc2

Figure 2.1: A SMILES code example: Pyrimethamine

Aromatic atoms are coded in lower case (c, n) with non-aromatic atoms are in upper case (C, N, Cl). Hydrogen atoms are not coded, but are assumed to occupy all remaining available bonding sites.

As an example, the chlorobenzene fragment is coded as a chlorine atom attached to an aromatic ring, which in turn is attached to another aromatic ring at the *para* location. The numbers '1' and '2' in the SMILES code mark the beginning and end of the first and second rings.



Clc2ccc(*rest_of_structure*)cc2
1 2 3 4 5 6

Figure 2.2: Encoding a SMILES code fragment

SMARTS (SMiles ARbitrary Target Specification)

SMARTS patterns are an extension of the SMILES notation, and can be used to describe substructures in the molecule. This allows precise substructural searches to be made in a compound database, and it follows that the exact matching of moieties (functional groups) will enable the clustering of compounds that are similar in such respects.

Two further fingerprints (FP3 and FP4) using 55 and 307 SMARTS patterns can be created in OpenBabel that can then be used to run substructure searches.

InChI strings (International Chemical Identifier)

Unlike SMILES codes, every structure has a unique InChI string, and it is possible to convey more structural information with their use. The structure is described in several layers which record the chemical structure, charges, stereochemistry, and isotope variations.

InChI keys

These keys are a condensed digital representation of InChI strings, introduced to improve how searches are handled for InChI strings. The keys are limited to a fixed length of 25 characters, with the first 14 characters used for the chemical structure, 9 representing the other description layers, and two for the version number and checksum. It is possible (but likely to be extremely rare) that two compounds can be represented by the same key.

InChI strings cannot be rebuilt from InChI keys, and therefore need to be fully linked to the appropriate records when used for searching.

MACCS keys

These are keys provide a 166 bit fingerprint for the structure, based on 166 fragments considered important in medicinal chemistry. Again, this fingerprint can subsequently be used for substructure similarity matching across candidate compounds.

2.8 Similarity searching using SMILES codes

Once an indicator compound has been identified, there is a need to find structurally similar molecules as these are more likely to exhibit similar properties than those chosen by stochastic processes. In order to achieve this, several techniques have been developed through that might be used to describe the similarity between SMILES strings e.g. Tanimoto similarity (**Tanimoto, 1957**), Dice coefficients (**Willett *et al*., 1998**), Monte Carlo simulations (**Toropov *et al*., 2009**) and digital compression (**Melville *et al*., 2007**); in some instances these techniques can also be applied to substructure similarity searches.

For this project, the Tanimoto coefficient was adopted as a similarity measurement to maintain a consistent approach to earlier research from which the prototype Active Learning algorithm was developed.

How can SMILES codes be used for similarity searching?

The coded SMILES representation can be split into fragments; this allows a fingerprint to be built of the active groups attached to the molecule. One such fingerprint, Daylight FP2 (**James *et al*., 1995**), is compiled using the following rules:

1. Linear small molecule fragments up to 7 atoms in length are recorded.
2. Single atom fragments of C, N, O are ignored.
3. The fragment is terminated when the atoms form a ring.
4. There is only one record of each fragment type, including those listed in reverse order.
5. Each fragment is numbered from 0 to 1020; this is used to identify it in a 1024 bit vector.

The fingerprint information extracted and collated in the 1024 bit vector for each molecule can then be usefully evaluated by comparing it to that of another molecule. One common method of comparison is the Tanimoto Similarity coefficient (**Tanimoto, 1957**), which is calculated from the number of ‘on’ bits of information in two objects that are similar, divided by the total number of occurrences:

		Object B		
		0	1	Total
Object A	0	d	b	$d+b$
	1	a	c	$a+c$
	Total	$a+d$	$b+c$	n

Figure 2.3: Comparison of encoded information for two objects

$$Tanimoto\ Similarity = \frac{c}{(a+b+c)} \quad (1)$$

Other methods defining the similarity between two fingerprints have been developed e.g. the Dice coefficient where the number of 'on' bits in the intersection of the two sets is divided by the average size of the features:

$$Dice\ coefficient = \frac{2c}{(a+b)} \quad (2)$$

The Dice equation is monotonic with Tanimoto, i.e. it will give the same ranking order of molecules based on the same fingerprints, but will return different values for the coefficients.

Other approaches take the number of 'off' bits into consideration, on the basis that these offer further information:

$$Matching\ coefficient = \frac{c+d}{(a+b+c+d)} \quad (3)$$

It has been argued that the maximum fragment size (7 atoms) represented by the standard Daylight 0/7 FP2 fingerprint is too low to represent some significant structures within the molecule, and Daylight 3/10 has been proposed as an alternative (**McGaughey *et al.*, 2007**) where fragments of 3 to 10 atoms are used. OpenBabel fingerprints are limited in this respect, but there are alternative sources of open source software e.g. RCDK (**Guha, 2007; Guha and Guha, 2013**) that allow suitable variants to be built**.**

SMILES fragmentation

SMILES codes have been used to find similarity between molecules by fragmentation, and subsequently predict relative chemical activity (**Bringmann and Karwath, 2004**; **Karwath and De Raedt, 2006**); such techniques yield simple, interpretable patterns, and can provide powerful results when used alongside other tools for extracting information from chemical databases. Fragmentation fingerprints will also have the advantage of smaller size, and this is beneficial when searching a large chemical space.

Similarity searching using the FP3 or FP4 fingerprints from the SMARTS patterns might be used alongside FP2 similarity searches; combining methods should give search results that are more relevant to potential druglike compounds, and these results could then be added to datasets for parameters depicted by the "Rule of 5".

Monte Carlo simulations for similarity searching

Monte Carlo simulations are iterative computational techniques that can be used in optimisation problems. Several attempts to build Quantitative Structure Activity Relationships (QSAR) using SMILES strings and standard data sets have been reported as promising:

For example, a standard set of $logTA100$ mutagenicity data for 48 nitrated polycyclic hydrocarbons (PAHs) has been used to develop a number of QSAR models (see section 2.10). One such technique based on Monte Carlo simulation employed an alternative use for SMILES codes (**Toropov *et al*., 2009**); this built on earlier work (**Toropov and Benfenati, 2004 & 2007; Toropov and Schultz, 2003**).

When investigating potential for similarity searching for this thesis, the Monte Carlo technique was found to be simple to describe and understand, and appears to give powerful results. However, there was no indication of its computational needs or the extent of the chemical space across which the model could be used. It would be difficult to predict how it might scale up to requirements for HTE. Also, prior work was with small, specific data sets so might have been susceptible to overfitting.

2.9 Large compound libraries for drug development

Pharmaceutical companies each hold their own collections of drug-like compounds, and it is estimated that more than 20 million diverse compounds now exist for development work. On a smaller and more accessible scale, commercially available libraries have also been built; these include the Maybridge Hitfinder Library and the ChemBridge collections. Commercial drug libraries are available in standard 384 well plates, and can be used to build application-specific daughter libraries.

Maybridge Hitfinder Library

This is a collection of 14,400 compounds, selected to provide a diverse range of pharmacophores (active fragments) in a relatively low number of candidates. The library closely follows Lipinski's "Rule of 5", and the full Maybridge collection of 53,000 compounds represents some 87% of the 400,000 pharmacophores estimated to be in existence, and the Hitfinder library has also been shown as highly diverse (**McGregor and Pallai, 1997**). The Maybridge Hitfinder library was used by Eve as its main source of chemicals; it was used to prove the robot's ability to independently discover drug activity, at a much lower cost than other alternatives (which might in turn have given a richer collection of hits).

ChemBridge MicroFormat Library

ChemBridge offer a number of libraries both for high throughput studies and for highly specific screens; their full list is in excess of 850,000 compounds. The ChemBridge MicroFormat library contains 180,000 components covering a wide chemical space.

Many other libraries exist, largely for specialised applications. In addition, interest in examining natural products is also growing (**Koehn and Carter, 2005**), leading to the development of such initiatives as the 5000 plant extracts compiled for the University of Strathclyde's Worldwide Natural Product Library (SIDR) **(Harvey *et al.*, 2010**). This is ostensibly similar in design to libraries of drug-like compounds (i.e. 96 well plates), but each well/plant extract might represent several active ingredients which in turn would need to be isolated in the case of hit identification.

2.10 Libraries of late clinical stage pharmaceuticals

The Comprehensive Medicinal Chemistry database listed 6683 compounds (**McGregor and Pallai, 1997**) that had been evaluated as therapeutic agents in humans, and was estimated to be growing by approximately 250 compounds *per annum*. Data associated with these compounds are available for *in silico* work, but their physical availability is a more difficult problem. There are ongoing attempts to build centralised libraries (see the examples below) containing all such compounds, but only about 40% are readily available. The others are largely compounds that have entered Phase II trials, but not commercialisation; to contribute to the library, they would need to be synthesised (following receipt of appropriate permissions).
The DrugBank initiative (**Knox *et al*., 2011**; **Wishart *et al*., 2006**) moves on another step by combining drug and target data; this is available in the public domain for *in silico* modelling (6811 compounds, linked to 4294 targets).

The Johns Hopkins University Clinical Compound Library (JHCCL)

The full JHCCL is the largest physical collection of approved compounds currently available for drug screening; to date, it consists of 3100 active ingredients found in known drugs (www.jhccsi.org/background.html). Various versions exist; the one made available for Eve contained 1600 FDA and non-USA-approved components in 7 plates, assembled in 384-well plate format carrying 5 mM stocks.

It has been mooted that the library be expanded to include all therapeutic agents known to clinical medicine (~11,000), with a view to using them as a screening tool against all neglected diseases (**Chong and Sullivan, 2007**).

The Prestwick Chemical Library

The library contains 1200 small molecules approved for drug use, selected based on known bioavailability and safety in an effort to minimise the number of potential hits with properties too borderline for further development stages.

The Spectrum Collection

This collection of 2000 compounds includes known drugs (50%), purified natural products (30%), and other bioactive compounds (20%). The drugs are largely out of patent, and have well understood properties. The natural products were chosen based on structural diversity but have largely unknown activity. The remaining bioactive products have known activity (e.g. for herbicide/pesticide uses) but have no therapeutic approval for humans.

The NCGC Pharmaceutical Collection (NPC)

This collection has resulted from collaboration between several USA government agencies, and includes 2400 small molecules approved for drug use in the USA (FDA), European Union (EMA), Japan (NHI) and Canada (HC). Further synthetic work is expected to build the library, and it is publically available for high-throughput experimentation (**Huang *et al*., 2011**).

Chemical library/source	Compounds	Status
Maybridge Hitfinder	14400	Available for Robot Eve
Johns Hopkins Clinical Compounds (JHCCL)	1187*	Available for Robot Eve
Full Maybridge library	~53000	Available for purchase
Full JHCCL	~3100	
All approved drugs	~3500	
All drugs (minimum Phase II trials)	~11000	
All synthesised drug-like	$\sim 2\times10^{7}$	
All possible drug-like	$>10^{50}$	

* 1187 discrete compounds remained in the Aberystwyth copy of the JHCCL after removal of duplicates, problem wells, and compound wells incompatible with Eve's liquid handling systems.

Table 2.2: Compound sources for use by Eve

2.11 Quantitative Structure Activity Relationships (QSAR)

QSARs have been used for many years to predict activity levels in biological and chemical systems; they take a performance dataset relating to the known structural parameters of a chemical and/or reaction mechanism, and use this data to predict the activity of other compounds under similar conditions. Larger and more complex groups of molecules and datasets can lead to highly fitted QSARs, which in turn require significant computing power to exploit.

The state of the art was reviewed in (**Dudek *et al*., 2006**). It is possible to build 2D QSARs based on several hundred attributes for a compound, describing topological, geometric, electrostatic and quantum-chemical properties across the range of possible fragments that might contribute to activity. Application of such models across a large compound library suggests the need for large computing capacity, although the literature more regularly deals with small datasets in highly specialised applications. Extending QSAR to 3D models involves a yet higher level of complexity, with a subsequent impact on the time taken to select candidate compounds. Adoption of a Machine Learning approach (see section 2.11) might allow simpler activity/descriptor relationships to be built from a training set, although this might limit the search space of the model.

Where more than one activity is being measured for an assay, advances in modelling and computation now allow separate models to be built and combined using multivariate QSARs (**Arodz and Dudek, 2007**). The possibility of extending QSAR predictions across a range of parasite species has been suggested (**Prado-Prado *et al*., 2010**). Both of these approaches used artificial neural networks, but the latter concluded that this method offered little improvement in precision over linear methods e.g. linear discriminant analysis (LDA).

The value of QSARs is reflected by the efforts shown in this research discipline (**Tropsha, 2010**). It is emphasised that good practice be maintained around data management, and that models need to be validated externally to avoid oft levelled criticism; the problems faced by QSARs to label structural outliers were discussed, suggesting that this challenge has yet to be met.

The tools used to build a dataset are not readily available to non-commercial organisations. The pharmaceutical companies have proprietary methods, only glimpses of which are occasionally described in the public domain. The original plan for the Robot Scientist Eve project was to collaborate with Pfizer for QSAR studies as well as Active Learning, but changes to the company's research and development outlook led to this option being withdrawn.

Few QSAR techniques are available in Open Source form:

The Bioclipse workbench includes a limited QSAR function (**Spjuth *et al*., 2007 & 2009**), and has recently been upgraded to work with the statistical analysis software, R (**Spjuth *et al*., 2013**).

An open source version of 3D QSAR, Open3DQSAR (**Tosco and Balle, 2011**), is aimed at ligand-based drug discovery, where structural/pharmacophore alignments are considered. Limited studies have been published using this software in comparison with existing commercial versions e.g. CoMFA, CoMSIA (**Ghasemi and Shiri, 2012**).

Several specialist software companies promote their tools as available to academics, but have limitations placed upon them, as might be expected of such collaborations.

2.12 Machine Learning and data mining

When running an HTE screen, it is imperative that there are no computational delays to the compound selection process. This is not a problem when simply mass screening a library, but is potentially a concern when a cherry-picking selection is required. If the assay is able to be screened rapidly, this will provide a limit to the time available for the compound selection method. The limits imposed by the yeast growth generated by Eve are not arduous in this respect, but later simulation work using alternative targets might be problematic if selection cycle times are long.

The application of an array of Machine Learning techniques for drug development are discussed in (**Barrett and Langdon, 2006**) and (**Dudek *et al*., 2006**), and it is noted that these techniques offer a means of dealing with the huge amount of data relating to problem solving in this field: “Many drug discovery problems can be expressed as the problem of finding a computer program”. One recurring concern in the literature is whether there had been sufficient industry take up of machine learning approaches, with the implication that pharmaceutical companies tending to stick conservatively to tried and trusted (albeit limiting) approaches.

Reducing the complexity of a problem in Machine Learning should lead to more straightforward computations. "A theory of the learnable" (**Valiant, 1984**) offers some insight into the importance of tightly specifying input information, to make computations feasible in a given time. If too many variables and classifications are involved then reaching any single conclusion becomes increasingly complex (a polynomial number of steps).

Activity prediction using QSAR can be conducted using supervised learning approaches, as the training set supplies output/activity levels for each item. Both linear and non-linear methods have been applied to such data, and the task is to learn the mapping of the input data to the output. Many elements of the input data also lend themselves to using unsupervised learning approaches through clustering strategies; the different methods of depicting structural similarities as shown in sections 2.7 and 2.8 can readily be exploited by such methods.

Input and output data sets can also be grown to specifically assist the learner; this Active Learning (section 2.13) is a form of semi-supervised learning

2.12.1 Supervised learning

Linear methods

The properties of each compound can readily be evaluated using linear statistical methods to develop QSARs. Linear models work well with small data sets of similar compounds, and are generally easy to interpret. Multiple Linear Regression (MLR) builds a model for the single response variable (i.e. activity) based on a linear function of the explanatory variables (i.e. compound properties) chosen to minimise the squares of the difference between predicted and measured activity (**McConway *et al*., 1999**). The technique is prone to over-fitting when explanatory variables are not truly independent, although it is possible to remove those with relatively low impact to provide simpler models. Partial Least Squares (PLS) aims to improve on MLR by reducing the model to those explanatory variables that are independent. Linear Discriminant Analysis (LDA) aims to reduce dimensionality for classification rules by transforming the data set to separate the resultant classes more effectively (**Alpaydin, 2004**) with minimal effect on in-class variance.

Non-linear methods

***k*-Nearest Neighbour (*k*NN)**

The activities of *k* known neighbouring compounds are used to predict that of the unknown (**Cover and Hart, 1967**). The effectiveness of this technique is potentially limited when nearby neighbours are sparse, and have reduced similarity across the locality; this is likely to be the case when the training set is small. There are alternative approaches to use when low numbers of seed examples are available:

(i) Provide a similarity boundary threshold at which the predictive capability is switched off, although this will result in some isolated unknown compounds remaining unlabelled.

(ii) Take the example offered by the single nearest neighbour to provide the activity prediction.

Decision Trees

Activity predictions are based on a series of rules against which the properties of the compound are examined. The rules are defined by multi-way cross validation of the training set, so that the full rule set provides the best fit to the observed activity levels. Each rule/node asks which classification/leaf a single property will lead to, with two branches coming from each node. The whole structure is a series of iterative nodes & branches, leading to a leaf at the end of each pathway.

With large numbers of properties measured or calculated for each compound, it is ultimately possible to build a large number of rules, which in turn has the likely consequence of over-fitting. The decision tree can subsequently be pruned by removing pathways or individual leaves to reduce the overall complexity of the rules, allowing general and readily interpretable sets.

Application of decision tree QSAR models allows simple rules to be built (**Suenderhauf *et al*., 2012**; **Lira *et al*., 2013**) that can make accurate predictions within small to medium size data sets; they can suffer from limitations relating to over-fitting and lack of robustness (**Hammann and Drewe, 2012**), but appeal for their computational simplicity and interpretability if constructed carefully.

Support Vector Machines (SVM)

QSARs based on SVM methods aim to create a hyperplane that separates compounds according to their classification (**Burbidge *et al*., 2001**). The training compounds at the boundary of the hyperplane define its location, and therefore become the support vectors; in practise, linear separation is unlikely to occur, and misclassified compounds will exist on either side of the hyperplane. Because the SVM only uses information near or at the separation boundary, other simpler methods can remove compounds that will be distant from this region, thereby removing some of the computational complexity. Once the location of the hyperplane is defined, predictions can be applied to unknown compounds.

Increasing the dimensionality of the data by feature mapping techniques (**Li *et al*., 2009**) can assist in realising the hyperplane if a simple low dimension approach gives insufficient discrimination.

2.12.2 Unsupervised learning

For unsupervised learning (**Ghahramani, 2004**), there is no defined training set with predetermined inputs and output values against which the model will learn; instead it uses these data to recognise patterns that form beyond what might be expected from general noise.

Clustering

Clustering uses similarities in the input data to group items together. Cluster size and usefulness will vary depending on the criteria by which they constructed; whilst knowledge of the value of these inputs might be useful, the ability of clustering to identify unexpected patterns is also of high importance.

Many clustering algorithms have been suggested; two of the earliest approaches (*k*-means and hierarchical) have been used for many empirical studies, and provide a solid base upon which to build. After splitting a data set into several discrete groups based on similarities in instances/data points, clustering can then be used as a starting point to apply supervised learning techniques (**Alpaydin, 2004**).

k-means clustering

The aim of *k*-means clustering is to find a pattern in the input data that gives *k* clusters whose components have minimised in-cluster variance (**Hartigan & Wong, 1979**); the clusters are then represented by their centre in subsequent analyses. One widely used example for *k*-means clustering is its application to customer segmentation in commerce, where similar customers are grouped so that their business can be exploited in similar fashion.

The size of the clusters is not controlled by this method, and it is possible to have a diverse range of cluster sizes; this can be useful, especially if there is interest in finding unrepresentative examples that might hold niche or rare information.

Hierarchical clustering

Simple strategies for hierarchical clustering are either divisive or agglomerative **(Jain et al., 1999)**. Divisive algorithms start with the full data set and divide it through a number of iterations until the desired level of detail has been reached. The

alternative agglomerative approach starts with clusters of size *n=1*, which can then merge based on local similarities to form larger groups (e.g. nearest neighbour clustering). Both methods typically use greedy selection, and both can be depicted in a tree-like structure.

2.13 Active Learning

The sheer quantity of data capable of being generated using High Throughput Experimentation is typical of much empirical work in our technological age. However, this seemingly unending supply of data needs to be filtered and categorised in order for it to have relevance, and this can be achieved to a certain extent using algorithms based on experience. Better still, rules derived from mathematical models and Machine Learning tools enable sense to be extracted, but there still remains the question of how to control the data generating process to make it more efficient.

The role of Active Learning is to make improvements to the model using fewer resources than might be required with simple supervised or unsupervised learning approaches; the training set is grown by the learner itself as it learns, and the next sets of data to be gathered are chosen on the basis of how their knowledge might improve the model. Active Learning can therefore be considered as a form of semi-supervised learning.

One aspect of the human learning process is the ability to adjust or evolve a query when additional experience is available. For Machine Learning, it may be seen that query techniques can be measured for effectiveness; if these can subsequently be honed either by adjusting the rules upon which the query is based, or by directionally improving the data patterns upon which their selections rely, these active changes can lead to a more efficient selection process.

If labelled examples are to be used to search the space in which the unlabelled examples exist, the queries should ideally be able to penetrate and explore all regions. Greedily searching for items that are most similar to existing examples will lead to an initial gain, but this is likely to rapidly fail as no significant new information is available to improve the model. Similarly, searching areas far away from known examples needs to be done in moderation. In general, an Active Learning model needs to be able to balance the regions it searches, and retain the ability to both explore and exploit the unknown space.

A general repository of Active Learning techniques is **(Settles, 2010)**; this online resource was being periodically revised after its initial appearance (2008), but does not seem to have been updated since 2010.

2.13.1 Active Learning in drug discovery

There seems to be little work reported on AL techniques in relation to drug discovery. It is considered likely that this is due to limited availability of assay data for development of AL algorithms, hence the potential advantages previously described for research using Robot Eve: the potential gains in efficiency of compound selection make this an area worthy of further investigation.

Work reported in (**Warmuth *et al.*, 2003**) aimed to find compounds in a large collection such that the fewest number of iterations of biochemical testing was conducted. Selection of compounds was conducted by exploiting the maximum margin hyperplane generated by Support Vector Machines.

The separating hyperplane allows different strategies to be used to choose the next compound for examination:

(i) Random.

(ii) The furthest on the positive side [exploitation strategy].

(iii) The closest to the previously known actives.

(iv) Those nearest to the decision boundary [exploration strategy].

Strategy (i) does not use earlier information, and increases the number of hits linearly; (ii) selects those most likely to be active, which finds many compounds in few iterations at the expense of better modelling of the problem; (iv) allows better modelling of the SAR; (iii) only searches locally, based on current knowledge, and will not find actives that are remotely located.

The Active Learning was carried out using descriptors generated by DuPont's in-house software (and using datasets supplied by DuPont Pharmaceuticals). It was also shown empirically that the activity of an 'active' compound is uncorrelated to its distance from the hyperplane. Alternative Machine Learning techniques (Voted

Perceptron, Bayes Point Machine) were discussed in (**Warmuth *et al*., 2001**); it was stated that they gave similar performance on this data to the SVM approach.

Hierarchical sampling for Active Learning (**Dasgupta and Hsu, 2008**) presents a potential technique for identifying the purity of data clusters during the learning process. Extrapolating the techniques therein:

By using a semi-supervised pre-clustering routine, and sampling within these clusters, it might be possible to identify spaces where the activity is unknown or uncertain in comparison to spaces where high certainty over activity or inactivity abounds.

2.13.2 *Active k-optimisation* strategy

One of the major goals for Eve's data analysis was to build algorithms to predict active compounds. The array of information contained in assay data includes raw data for yeast target growth profiles, labelled classifications for activity, toxicity etc., and structural representations of the compounds.

The prototype method for this work was to be based on the *active k-optimisation* strategy (**De Grave *et al*., 2008a**). This strategy is introduced for machine learning as a way of finding and ranking the *k* best alternatives for evaluation, using Gaussian process to provide a mechanism for developing this model.

The general idea is to develop a process to find more than one target (other than the optimal solution) to take into the next step of HTE. The work is based on having a finite library of examples, of which the results from the known ones can be used to pick the best unknowns for evaluation.

In this particular approach, the goal is to pick targets that have the best chance of success (the **maximum predicted** strategy) for comparison to several other existing approaches (**King *et al*., 2004; Vandezande *et al*., 2005; Jones *et al*., 1998**). The lower confidence bound criterion (**optimistic**) (**Cox and John, 1997**), selecting the sample with the highest probability of improving the current solution (**most probable improvement, MPI**) and efficient global optimisation (EGO) to give **maximum expected improvement**.

Specific application of the *active k-optimisation strategy* to the drug screening process is provided in (**De Grave *et al*., 2008**); this describes the analysis of the NCI60 dataset (US National Cancer Institute, 60 anticancer drug screen). The full techniques upon which this strategy is based are given in (**Bishop *et al*., 2006**).

2.13.3. Transfer Learning

It is expected that implementation of a Machine Learning process will be based on a set of data split into training and validation examples, together with an independent test set. However, it is possible to provide additional strength to the training data by using knowledge gained from previous processes on different examples.

The main questions are: what to transfer, how to transfer, and when to transfer. The value of the transferable knowledge needs to be pre-determined, in order to avoid a negative transfer effect (**Pan and Yang, 2010**).

In the context of a QSAR Active Learning algorithm, the knowledge transferred is embedded in a set of seed compounds previously found to be active against other targets. There is no guarantee that similar activity will be found with the new target, so the simplest approach might be that an active compound set is retrained in the new experiment. This is an application of inductive transfer learning, which in turn shows some similarities to multitask learning (**Caruana, 1997**).

For Eve, knowledge of the phylogenetic similarity between target parasites (**Tibayrenc *et al*., 1990**) might be an additional tool to aid selection of training data for new experiments.

2.13.4 Rare category/class detection

Various models are employed to detect rare categories/events, in an effort to isolate useful anomalies. It has been found that the most successful approaches include subjective classification by the user (i.e. a “hunch”!!); this was suggested alongside a novel approach (**Pelleg and Moore, 2004**) which used a known component applied to a data set whose contents are then ranked by a probability density function. The untested examples with highest rankings (i.e. most anomalous) are classified and fed back into the next iteration. The authors admit this is an intuitive/empirical approach that worked with their large database (> 10^6 records), without offering an

explanation; in my opinion, it is feasible that large outliers are more likely to be interesting as, if the data is in a Gaussian distribution, lower numbers of noisy points are expected at the extremities.

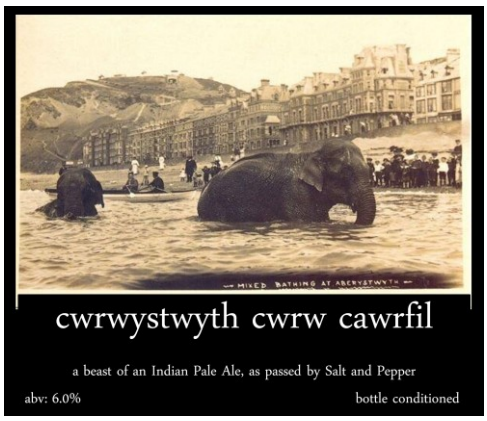
The potential for identifying knowledge gaps (see 2.11.2) also lends itself to rare class detection and mining. The methods identified in (**Han *et al*., 2009**) suggest that the distribution of the data set will influence the techniques required to identify rare classes.

Chapter 3

Development of intelligent robotic systems for drug discovery

Adventures with yeast, part 3 of 7: Cwrw Cawrfil

3.8 kg	*Maris Otter pale malt*
1.0 kg	*torrified wheat*
0.7 kg	*white sugar*
100 grams	*East Kent Goldings hops*
100 grams	*Bobek (Styrian Goldings) hops*
1 packet	*Safbrew T58 wheat beer yeast*
1	*Irish moss tablet*



Add the grains to the mash tun with 20 litres of water. Steep at 70-75°C for two hours. Remove the grain sack and allow it to drain into the wort; sparge it with boiling water until the sugars have been depleted. Remove 500 ml of wort and use it to make a yeast starter. Meanwhile, add the white sugar to the bulk wort, and bring to the boil; maintain the volume at approximately 25 litres. Add 60 grams of each hop when the wort is at a rolling boil, and boil for 60 minutes; add 20 grams of each at 15 minutes from the end of the boil, and a further 20 grams of each at 5 minutes from the end, together with the Irish moss tablet.

Cool the finished wort; decant it to a fermentation vessel and add the yeast starter. Record the original gravity (~1058), and the final gravity (~1014 after 2 weeks). Prime the finished brew with 100 grams of white sugar in 1 litre of boiled/cooled water prior to bottling.

3.1 Robot Scientist Eve project overview

Robot Scientist Eve has been designed to screen compound libraries *en masse* against chosen targets, and to switch from this mass screening mode to a cherry-picking mode. The latter allows Eve to examine activity in more detail across a wider range of compound concentrations.

Eve's output has been used as a primary data source for this thesis. The physical screening processes were being developed concurrently with this thesis' work on drug-like activity measurements. Significant effort was put into building robust data sets that would underpin later work on drug discovery Active Learning approaches.

This chapter covers the fundamental building blocks of the project, the transformation processes for the raw data provided by the screens, and example results from the screens conducted whilst Eve was based in Aberystwyth.

The relative activity of compounds and negative controls was used to categorise potential hits, separating them from possible toxic compounds, inactive ones, and noise. The use of two parasite target strains and one control strain (human) in each assay meant that the data analysis steps could also make use of intra-well strain performance.

The main data corrections were based on negative controls, for which repeatability has been shown; it would have been advantageous to have more positive control data too, but the nature of such items (known protein-specific interactions) meant that this was near impossible to arrange. Components from the data set were used to build a clear set of rules that displayed a good fit to the observations, and made sense in terms of the growth of the yeast.

The confirmation screen data were also categorised successfully in terms of activity against the given target, with an indicator for possible toxicity. The rules for the confirmation screens were based on a comparison of the shape of the concentration/activity curve of the parasite strain versus the human strain.

Eve's compound libraries have been tested against several parasite strains, and many active compounds have been confirmed, including some from the library of

existing drug therapies (JHCCL); *in vivo* experiments were conducted on some of these compounds, and these results are included later in this chapter.

An alternative approach to identifying potentially active compounds was provided by the School of Biological Sciences, University of Cambridge; its performance was compared to the methods mentioned above.

3.1.1 Drug-like compound libraries

Eve has been designed to grow multiple strains of modified yeast in the presence of drug-like compounds selected from chemical libraries; typically, two parasite strains were used to make up the assay for a single experiment, with a human strain as a control. The relative growth rates of the yeast strains were then used to determine the activity of the compounds against the genetic modification. The building blocks for these experiments were the chemical compound libraries (Table 3.1) and the yeast strains (Table 2.1):

Chemical library/source	**Compounds**	**Status**
Maybridge Hitfinder	14400	Available for Robot Eve
Johns Hopkins Clinical Compound Library (JHCCL)	1187*	Available for Robot Eve
Full Maybridge	~56000	Available for purchase
Full JHCCL	~3100	
All approved drugs	~11000	
All synthesised drug-like	$\sim 2\times10^{7}$	
All possible drug-like	$>10^{50}$	

* After removal of replicates and compounds incompatible with Eve's liquid handling

Table 3.1: Compound libraries for Eve

3.1.2 Hardware and software

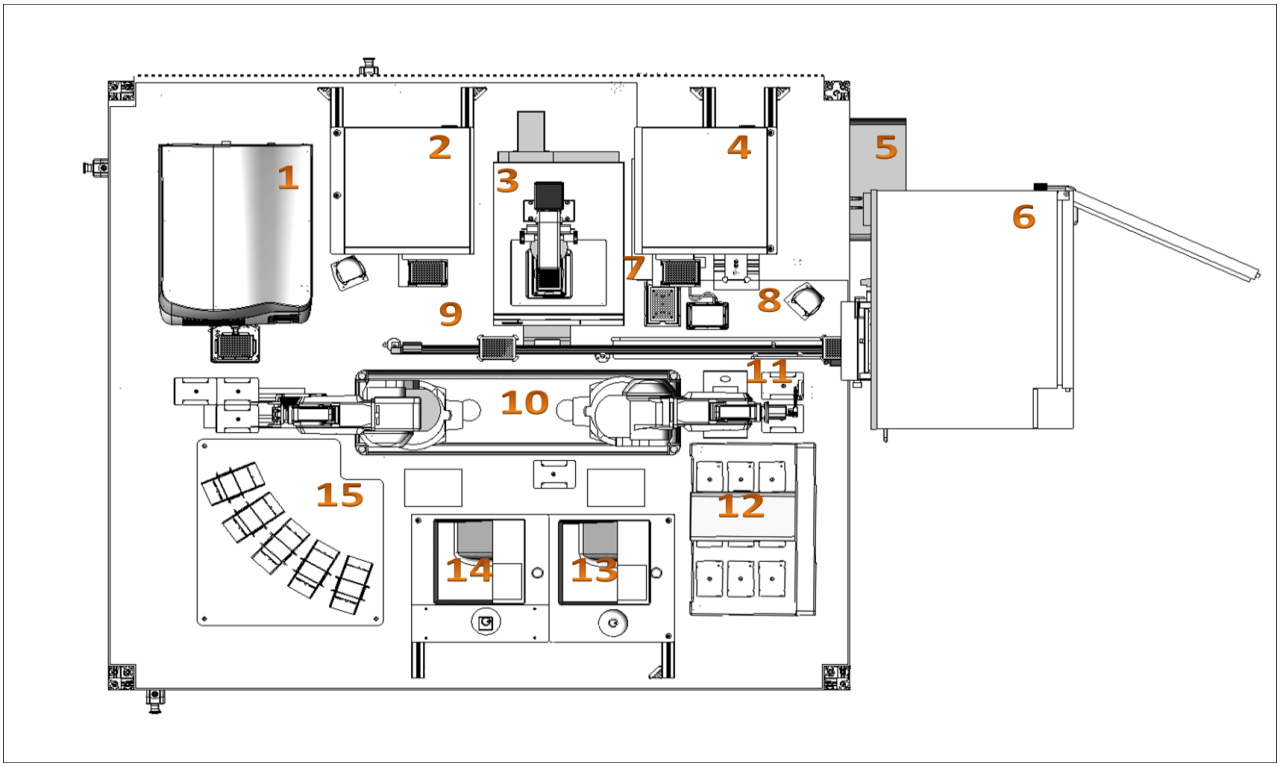
Eve was designed as a fully integrated laboratory robotic system (**Sparkes *et al.,* 2010**); a variety of component instruments were included to enable future flexibility when developing alternative experiment regimes.

The main components were:

- a dry store for holding chemical compounds,
- an incubator,
- liquid handlers and multidrops for transferring materials between libraries and experiment plates,
- shakers,
- a capper-recapper,
- plate readers for measuring growth across a wide range of the UV-visible spectrum,
- imagers for recording cell growth and morphology.

The robot control software has been written to execute mass and cherry-pick screens, and generate data sets based on the resultant yeast logistic growth curve. In addition, there was a requirement for data analysis and intelligent compound selection processes; these were to form part of the work for this thesis.

Figure 3.1: Profile photograph of Robot Scientist Eve



	Equipment
1	Labcyte Echo 550 acoustic liquid handler
2	BMG Pherastar reader
3	MDS ImageXpress Micro cellular imager
4	BMG Polarstar reader
5	Cytomat 2C435 incubator
6	Cytomat 6003 dry store
7	FluidX DC-96pro capper/recapper
8	Variomag teleshake plate shakers and Metrologic Orbit 1D barcode readers
9	Cytomat linear actuator track
10	Robot plinth holding Mitsubishi robot arms; models RV-3SJB and RV-3SJ
11	FluidX Xtr-96 tube rack 2D barcode scanner
12	Agilent (Velocity 11) Bravo liquid handler
13	Thermo Combi-nL multidrop
14	Thermo Combi multidrops
15	Consumables stacks for microplates, tube racks and tips

Figure 3.2: Schematic layout for Eve (plan view)

3.1.3 Screening plates and growth curves

General screens were made up of a number of 384 well plates, with each well containing three yeast strains for the chosen assay. 320 wells per plate were dosed with 50 nl of compound (10 μM) dissolved in dimethylsulphoxide (DMSO), with 64 wells used for controls; the plates contained up to 20 positive controls (5 separate compounds) and 44 negative controls (triple strain assay + DMSO), or contained 64 negative controls.

In vitro assays using HTS have typically used compound concentrations in the range of 1 to 50 μM (**Keser and Makara, 2006**); the concentrations used by Eve were based on these practices:

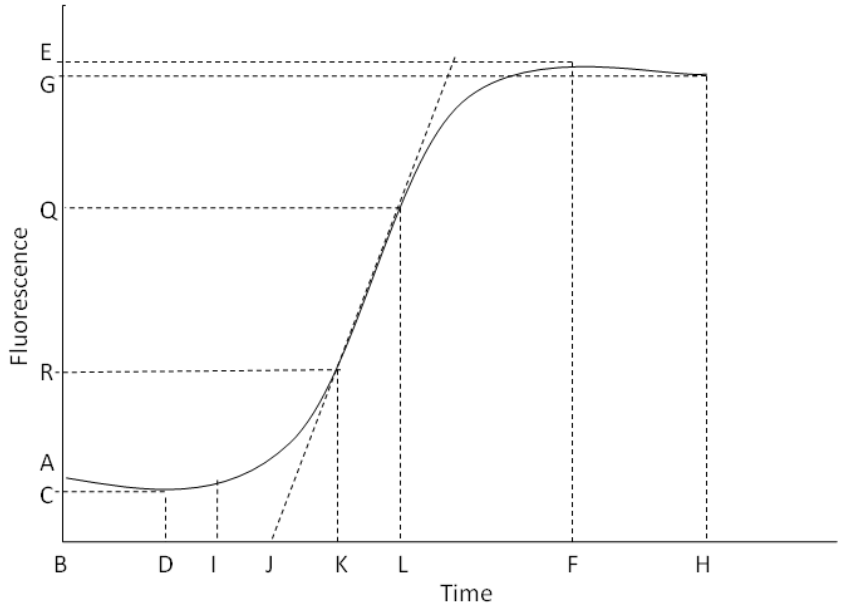
- 10 μM is a standardised concentration for screening purposes
- The confirmation range (0.5 to 10 or 20 μM) was kept broad but tended away from higher concentrations (unrepresentative of therapeutic doses).

Originally, plates for cherry-pick and confirmation screens contained eight replicates of eight compounds, at 6 concentrations (0, 1, 2.5, 5, 10, 20), thereby also having 64 negative controls by default. Later confirmation screens contained four replicates of sixteen compounds, at 6 concentrations (0, 0.5, 1, 2.5, 5, 10 μM); see section 3.2.5 for more details.

The plates were incubated for 40 hours, and growth was recorded every 90 minutes through measurement of fluorescence at the three strain emission wavelengths.

Logistic growth curves were then derived, and parameters A to P (see Figure 3.3) calculated **(King *et al.,* 2009).**

The full list of assays is given in Appendix A.1; this lists the protein target and parasite combinations for each triple strain experiment.



A	startvalue	I	lagtime
B	startvaluetime	J	miylagtime: a biologist's popular estimation of the lagtime, as suggested by Professor Mike Young (MIY).
C	minvalue	K	startlinear
D	minvaluetime	L	endlinear
E	maxvalue	M	linearslope = (Q - R)/(L - K)
F	maxvaluetime	N	doubletime = 1/M
G	endvalue	O	durlinear = L - K
H	endvaluetime	P	snratio (signal/noise ratio)

Figure 3.3: Diagram of a typical logistic growth curve

3.2 Development of data analysis pipelines for active compound identification and confirmation

Eve's data output for a mass screen contains the parameters listed in Figure 3.3 for each yeast strain, together with labels for sample identification purposes. The first task was to identify which of these parameters could be used to identify activity of a compound against the chosen target. The criteria eventually selected for routine analyses are listed in Appendix C.1. There was also a strong requirement to show consistency between measurements in discrete screens. The need for Eve to provide repeatable results, especially when operating in confirmation/cherry-picking mode, is of fundamental importance when running objective studies.

The data set for the initial triple strain mass screen (MS_63_1_15_20110414203535; see section 3.3.1) was used to identify potential active compounds; the three yeast strains were HsDHFR, PvDHFR and PfRdhfr. The data set was first checked for consistency through statistical analysis of the negative controls, and then abnormalities were identified versus these background controls. Visual classification of these abnormal growth curves allowed rules to be built (decision tree analysis) that could identify activity through using a wider set of the growth curve parameters.

3.2.1 Statistical analysis of negative control data

The results from the negative control wells provide the baseline against which candidate wells were compared. The negative controls (between 44 and 64 on each plate) only contained the three yeast strains in the growth media, together with DMSO (the solvent used for all candidate compounds). In-plate and in-batch repeatability of negative control growth rates could be used to verify full data sets.

With the exception of the 'start' and 'end' timings (Figure 3.3, B & H), it was expected that any of the logistic growth curve parameters might be affected by the presence of active compounds. The negative control wells were available for a baseline comparison on a plate-by-plate basis.

The doubling time (DT) parameter was chosen to determine:

i. Repeatability in-plate using the negative controls.
ii. Repeatability within a batch of plates using the negative and positive controls.

iii. Repeatability across a full screen of 45 plates using negative and positive controls.

iv. Compounds that significantly affect growth in comparison to negative controls.

The mean and variance of the DT results were calculated for the DHFR-TS3 assay negative controls with the assumption that they fit a normal distribution (under the central limit theorem):

Batch	Plates	mcherry DT HsDHFR		sapphire DT PvDHFR		venus DT PfRDHFR		Sample size
		Mean	Variance	Mean	Variance	Mean	Variance	
1	2125 to 32	2.05	0.0046	3.64	0.0062	2.90	0.0037	415
2	2133 to 40	2.04	0.0062	3.52	0.0042	2.86	0.0024	514
3	2141 to 48	2.04	0.0077	3.15	0.0032	2.87	0.0035	514
4	2149 to 56	2.76	0.0560	3.80	0.0181	3.65	0.0281	514
5	2158 to 65	2.05	0.0055	3.39	0.0038	3.02	0.0036	440
6	2170 to 74	2.09	0.0037	3.56	0.0035	2.91	0.0024	298
All		2.19	0.0937	3.51	0.0531	3.05	0.0969	2695
Not batch 4		2.05	0.0061	3.44	0.0358	2.91	0.0065	2181

Table 3.2: Descriptive statistics of negative control doubling times

The results in Table 3.2 suggest that a problem arose when running batch 4. The individual plates for batches 1 and 4 were examined:

Plate	mcherry DT HsDHFR		sapphire DT PvDHFR		venus DT PfRDHFR		Sample size
	Mean	Variance	Mean	Variance	Mean	Variance	
2125	2.06	0.0030	3.69	0.0052	2.89	0.0021	44
2126	2.05	0.0042	3.67	0.0039	2.84	0.0023	44
2127	2.04	0.0066	3.65	0.0060	2.89	0.0028	45
2128	2.04	0.0049	3.64	0.0105	2.87	0.0039	45
2129	2.03	0.0043	3.64	0.0040	2.90	0.0032	44
2130	2.07	0.0062	3.66	0.0061	2.92	0.0038	65
2131	2.04	0.0037	3.60	0.0041	2.95	0.0019	64
2132	2.05	0.0036	3.62	0.0057	2.92	0.0029	64
2149	2.79	0.0613	3.77	0.0133	3.67	0.0427	66
2150	2.70	0.0371	3.76	0.0107	3.63	0.0222	63
2151	2.74	0.0552	3.79	0.0119	3.66	0.0278	65
2152	2.74	0.0638	3.77	0.0179	3.65	0.0265	64
2153	2.75	0.0897	3.79	0.0290	3.66	0.0392	64
2154	2.79	0.0458	3.82	0.0170	3.68	0.0193	64
2155	2.80	0.0539	3.86	0.0230	3.70	0.0236	64
2156	2.78	0.0385	3.85	0.0144	3.60	0.0193	64

Table 3.3: Descriptive statistics for negative control plates, batches 1 & 4

Just prior to commencing analysis of the above, it was identified that the yeast strains for batch 4 had problems in the pre-inoculation phase due to incorrect make up of the growth media, before use by Eve. The curves from batch 4 were not used in subsequent definitions of decision tree.

A two sample t-test (unequal sample size & variance) using the mean and variance for batches 1 and 4 gave the following test statistics:

	mcherry DT	sapphire DT	venus DT
t statistic	65.3	22.4	94.5
Degrees of freedom	616	851	672

Table 3.4: Population comparison, batches 1 and 4

This analysis strongly indicates a difference between batches 1 and 4 ($p<0.0001$), in agreement with the observation concerning a problem with the growth media.

The distribution of the negative controls DT for each strain in batch 1 was examined to make sure that they followed a normal distribution:

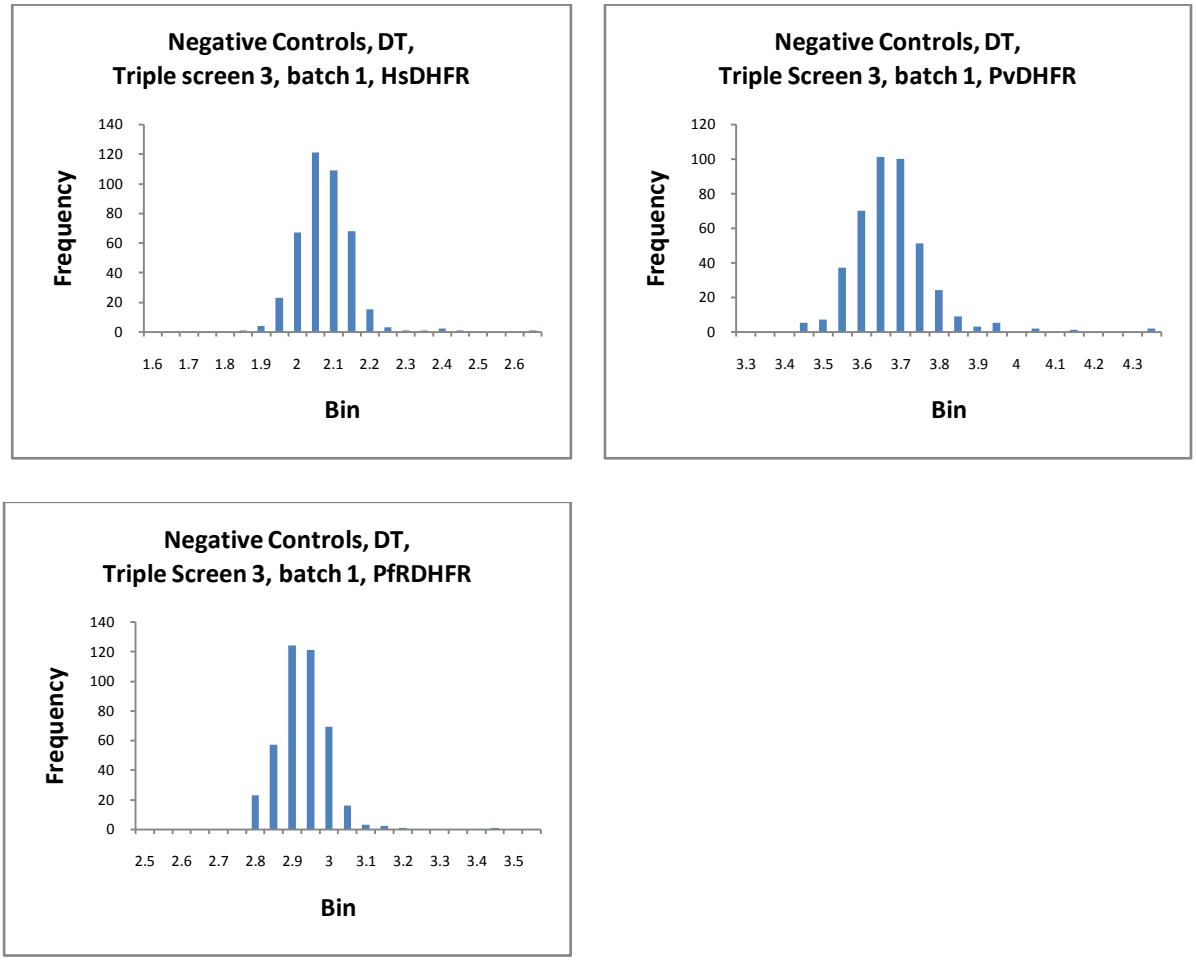


Figure 3.4: Population distribution for negative controls in batch 1 (plates 2125 to 2132) for all three modified yeast strains

3.2.2 Manual categorisation of triple strain screens

The DT results for the compounds were screened against the distribution of negative controls, with any result more than 2 standard deviations above the mean being flagged as potentially active. The aim of this step was to initially identify a batch of compounds with growth curves that differed from the norm, before conducting a deeper analysis of the reasons behind these differences; this process identified approximately 600 compounds. It is expected that 2.28% of compounds would naturally meet this criteria if the negative control and compound populations are Gaussian and similar; this is equivalent to 328 of the 14400 Maybridge Hitfinder library compounds.

Variation in DT was not expected to be the only measure of identifying activity, as it is possible for an assay to double at the same rate as a negative control but for a shorter period; it would therefore be limiting to select by this rule alone, as a compound so described would mistakenly not be identified as active.

The potential actives were then examined visually in order to gain a better understanding of why they had been flagged as abnormal; they were then classified according to the shape of the curves for each strain:

i. Some compounds autofluoresce at the same wavelength as the strains, thereby interfering with effective assessment; three curve shapes were identified having:
    a. High fluorescence throughout the run.
    b. High starting fluorescence, falling to meet the typical curve later in the experiment.
    c. Low starting fluorescence, but noticeably above the negative control; then generally staying above the typical curve.

ii. Toxic compounds were classified as those with lower intensity curves for all three strains; these were of three types:
    a. Strains with typical growth curves, but at a lower intensity.
    b. Strains showing little or no growth.
    c. Strains with long lag times, only beginning to grow late in the experiment.

Active compounds were visually classified according to their effect on the strain:

a. One or more strains with growth curves noticeably weaker (i.e. end-of-test growth <90% than the negative control), with one or more strains having normal growth curves.
b. One or more strains with growth curves noticeably weaker, with one or more strains having stronger growth curves.

iii. Other effects were sometimes seen, overlaying the regular growth curve:

a. The curves are initially as normal, but the signal decayed later in the experiment to give a curve with a sharkfin appearance.
b. Curves initially build as normal, but then continue to show a low growth rate when they should normally have reached a plateau.

Figure 3.5 represents the main classes of logistic growth curves produced by Robot Eve:

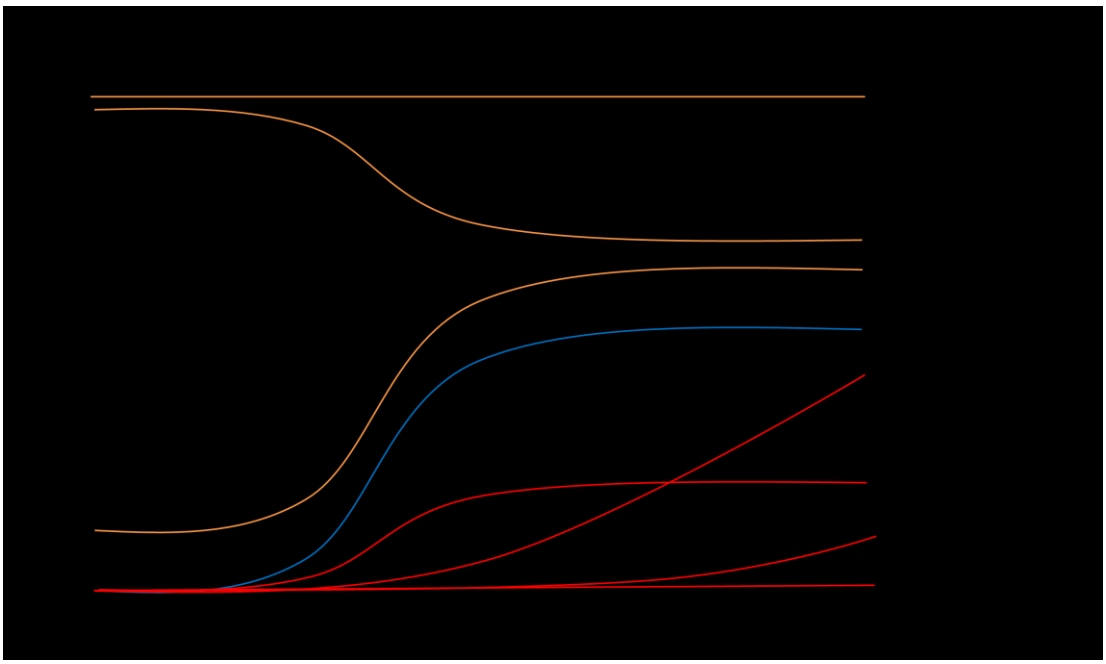


Figure 3.5: Typical growth curves seen in Eve experiments

Of the 600 compounds with an abnormal DT, 325 were categorised:

- 20 with strong and 9 with weak activity against HsDHFR
- 57 strongly and 64 weakly active against PvDHFR
- 9 active against PfRdhfr
- 83 toxic compounds (active against all three strains)
- 16 autofluorescent compounds
- 67 normal growth curves (inactive against all three strains)

The active results for PfRdhfr were generally of lower quality/certainty than those for the Hs/PvDHFR curves.

The above process for identifying candidate compounds was extremely labour intensive, but a necessary first step to produce learning/identification algorithms. This process also gave the author a better understanding of differences in growth patterns of yeast targets, and served as a visual check for Eve's behaviour.

Normalisation of output data

Prior to running machine learning routines, the well data for compounds needed to be normalised against the negative control data. The recently gained experience of variants in the shape of the growth curves suggested that comparisons of the end-of-test fluorescence would be of primary significance, together with initial, minimum & maximum fluorescence, and lagtimes. Fluorescence measurements were recalculated as ratios against the negative control, and lagtimes were treated as a difference to the negative control.

Total growth ratios for each strain were calculated using the fluorescence data:

$$strain\ growth\ ratio = \frac{end_{\ compound} - start_{\ compound}}{end_{\ neg\ .ctrl} - start_{\ neg\ .ctrl}} \qquad (4)$$

$$= \frac{G_{compound} - A_{compound}}{G_{neg\ .ctrl} - A_{neg\ .ctrl}}$$

Start values were calculated as a proportion of the of the negative control growth:

$$strain\ start\ value = \frac{start_{\ neg\ .ctrl} - start_{\ compound}}{end_{\ neg\ .ctrl} - start_{\ neg\ .ctrl}} \qquad (5)$$

$$= \frac{A_{neg\ .ctrl} - A_{compound}}{G_{neg\ .ctrl} - A_{neg\ .ctrl}}$$

Lagtimes and miylagtimes were calculated as a difference to the negative control:

$$strain\ miylagtime = miylagtime_{neg\ .ctrl} - miylagtime_{compound} \qquad (6)$$

$$= J_{neg\ .ctrl} - J_{compound}$$

Doubling times were recalculated as a direct ratio to the negative control:

$$Strain\ DT = \frac{DT_{compound}}{DT_{neg\ .ctrl}} \qquad (7)$$

$$= \frac{N_{compound}}{N_{neg\ .ctrl}}$$

3.2.3 Decision tree analysis of the first mass screen data set

A total of 81 discrete attributes were available across all three strains; these were tested against the categorised growth curves for 325 compounds. After systematic trial using Weka 3.6.2, C4.5 (J48) decision trees (**Witten and Frank, 2005)** on all attributes, four attributes for each strain gave rules that were readily interpretable:

1. End-of-test growth ratio.
2. Start value.
3. Doubling time.
4. miylagtime (point J on the logistic growth curve; miy are the initials of Professor Mike Young, IBERS, who has suggested this as the biologist's popular estimation of the lag phase (**King *et al.*, 2009**)).

The combined growth (sum of end-of-test growth for all three strains) was also used as an attribute in the final set of decision trees.

Rules were built using Weka 3.6.2, C4.5 (J48) decision trees with 10-fold cross validation (those with the lowest average error on the validation set; 10 training activities: train on $^{9}/_{10}$ of data and test $^{1}/_{10}$; superfluous attributes pruned):

Category	Rule	General interpretation/ comments
Fluoro	sapphire_startvalue > -0.080 [306 inactive instances]	No fluorescence interference if initial value is less than 8% above that of the negative control.
	sapphire_startvalue ≤ -0.080 sapphire_ratio ≤ 0.85 [11 active instances]	
	sapphire_startvalue ≤ -0.080 sapphire_ratio > 0.85 venus_startvalue ≤ -0.0007 [2 inactive]	
	sapphire_startvalue ≤ -0.080 sapphire_ratio > 0.85 venus_startvalue > -0.0007 [5 active/1 inactive]	

Table 3.5: Decision tree rules for fluorescent compounds

Category	Rule	General interpretation/ comments
Toxic	total ≤ 1.70 mcherry_ratio ≤ 0.91 [57 active instances] mcherry_ratio > 0.91 [2 inactive/1 active]	Summed ratio for the three assays is very low, and Hs (mcherry) growth is reduced.
	total > 1.70 venus_miylagtime ≤ -10.5 1.20 ≥ mcherry_ratio > 1.58 [17 active] 1.20 ≤ mcherry_ratio < 1.58 [3 inactive/1 active]	A long lagtime is an indicator of activity.
	total > 1.70 venus_miylagtime > -10.5 venus_ratio ≤ 0.76 0.71 ≤ sapphire_ratio ≤ 0.80 [5 active/1 inactive] 0.71 > sapphire_ratio > 0.80 [22 inactive]	Multiple low individual strain ratios as indicators of activity.
	total > 1.70 venus_miylagtime > -10.5 venus_ratio > 0.76 [214 inactive/2 active]	Negative examples when one of the strains has growth within 24% of the negative control; this reduces the likelihood of overall toxicity.

Table 3.6: Decision tree rules for toxic compounds

Category	Rule	General interpretation/ comments
Hs & toxic	mcherry_ratio ≤ 0.68 [85 active/6 inactive]	Low growth.
	mcherry_ratio > 0.68 mcherry_miylagtime ≤ -8.8 [11 active]	Long lagtime despite reasonable growth.
	0.68 < mcherry_ratio ≤ 0.79 mcherry_miylagtime > -8.8 [6 active/7 inactive]	Borderline growth (< 80% of negative controls).
	mcherry ratio > 0.79 [10 active/184 inactive]	Negative examples when growth within 21% of negative controls; still a few active curves at this level.

Table 3.7: Decision tree rules HsDHFR-active compounds

Category	Rule	General interpretation/ comments
Pv & toxic	sapphire_ratio ≤ 0.83 [186 active/16 inactive]	Low growth.
	sapphire_ratio > 0.83 sapphire_miylagtime ≤ -3.71 sapphire_DT ≤ 1.25 [2 inactive]	Long lagtime but reasonable growth rate.
	sapphire_ratio > 0.83 sapphire_miylagtime ≤ -3.71 sapphire_DT > 1.25 [7 active/1 inactive]	Long lagtime and slow growth rate.
	sapphire_ratio > 0.83 sapphire_miylagtime > -3.71 [86 inactive/11 active]	Negative examples, but still a few active curves.

Table 3.8: Decision tree rules PvDHFR-active compounds

After using Weka to provide classification rules, it became clear that the growth ratio, miylagtime, and doublingtime were of most relevance. Further decision trees were built for HsDHFR and PvDHFR data using only pairs of these attributes:

Category	Rule	General interpretation/ comments
(1) Rules using only growth ratio and miylagtime	sapphire_ratio $\leq$ 0.83 [186 active/16 inactive]	Low growth
	sapphire_ratio > 0.83 sapphire_miylagtime $\leq$ -3.71 [7 active/3 inactive]	Long lagtime despite reasonable growth
	sapphire_ratio > 0.83 sapphire_miylagtime > -3.71 [86 inactive/11 active]	Negative examples, but still a few active curves.
(2) Rules using only growth ratio and DT	sapphire_ratio $\leq$ 0.83 [186 active/16 inactive]	Low growth
	sapphire_ratio > 0.83 sapphire_DT > 1.47 [6 active/3 inactive]	Slow doubling time (growth rate)
	sapphire_ratio > 0.83 sapphire_DT $\leq$ 1.47 [12 active/86 inactive]	Negative examples, but still a few active curves.

Table 3.9: Rules using (1) growth ratio and miylagtime, and (2) growth ratio and doublingtime, for PvDHFR-active compounds

Category	Rule	General interpretation/ comments
(1) Rules using only growth ratio and miylagtime	mcherry_ratio ≤ 0.68 [85 active/6 inactive]	Low growth
	mcherry_ratio > 0.68 mcherry_miylagtime ≤ -8.8 [11 active]	Long lagtime despite reasonable growth
	mcherry_ratio > 0.83 mcherry_miylagtime > -8.8 [191 inactive/16 active]	Negative examples, but still a few active curves.
(2) Rules using only growth ratio and DT	mcherry_ratio ≤ 0.68 [85 active/6 inactive]	Low growth
	0.68 > mcherry_ratio ≤ 0.72 [3 inactive]	
	0.68 > mcherry_ratio ≤ 0.72 mcherry_DT ≥ 1.32 [6 active]	Slow doubling time (growth rate)
	mcherry_ratio > 0.79 [21 active/186 inactive]	Negative examples, but still a few active curves.

Table 3.10: Rules using (1) growth ratio and miylagtime and (2) growth ratio and doublingtime, for HsDHFR-active compounds

The rules obtained from the Weka decision tree analysis were still seen as specific to the individual TS3 strains and the Maybridge Hitfinder Library. The rules for each of the filters were then generalised to give a set using startvalues, growth ratios, doubling times and lagtimes.

Category	Rule	General interpretation/ comments
Fluoro	filter_startvalue $\leq$ -0.08	Compound is autofluorescent if the initial reading is >8% higher than the negative control.
Active for filter	filter_ratio $\leq$ 0.8	Low growth at end of test, <80% of the negative control.
	filter_ratio > 0.8 filter_DT > 1.5	Reasonable growth, but doubling time is >50% higher than the negative control.
	filter_ratio > 0.8 filter_miylagtime < -4	Reasonable growth, but lagged by >4 hours compared to negative control.

Table 3.11: Generalised categorisation rules for active compounds

3.2.4 Labelling the activity of compounds in a mass screen

The rules for this process were found using J48 decision trees in Weka (**Witten and Frank, 2005)**.

Attributes required:

F1 Fluorescence at start of test
F2 Fluorescence at end of test
MIY "Mike Young" lagtime
DT Doubling time

The attributes for the individual compound wells were evaluated against the respective mean of the negative controls (as calculated on a plate-by-plate basis). Compounds were filtered and labelled to describe their relative activity.

Filters:

A compound is labelled as ***autofluorescent*** if F1 is more than 8% through the range for the negative control:

$$\frac{F1_{compound} - F1_{neg.ctrl}}{F2_{neg.ctrl} - F1_{neg.ctrl}} > 0.08 \qquad (8)$$

A compound is labelled as a ***potential hit*** if the end-of-test fluorescence is less than 80% of the negative control:

$$\frac{F2_{compound} - F1_{compound}}{F2_{neg.ctrl} - F1_{neg.ctrl}} > 0.80 \qquad (9)$$

If the end-of-test fluorescence is more than 80% of the negative control, a compound may still be labelled as ***possibly active*** if:

(a) the lagtime is more than 4 hours behind the negative control:

$$MIY_{compound} - MIY_{neg.ctrl} > 4 \qquad (10)$$

or (b) the doubling time is more than 50% above the negative control:

$$\frac{DT_{compound}}{DT_{neg.ctrl}} > 1.5 \qquad (11)$$

The labelling process:

1. Label and remove the ***autofluorescent*** compounds.
2. If compounds are ***potential hits*** against all fluorophores (mcherry, sapphire, venus) label them as ***definitely toxic***.
3. Scoring a ***potential hit*** as 2, and a ***possibly active*** as 1, compounds not identified as ***definitely toxic*** are labelled as ***probably toxic*** if they score 5 or 6 across the three channels.
4. After removal of toxic compounds, active compounds can then be ranked either directly by using their end of test fluorescence ratio, or by taking the ratio of this value against the HsDHFR ratio.
5. All other compounds are considered ***inactive***.
6. The compounds labelled in the toxicity categories are probably worth examining further, as the activity of the test strain might be sufficiently different to the Hs strain to warrant a deeper investigation. Later confirmation work suggested that some toxic compounds might be of interest if examined at lower concentration.

3.2.5 Decision tree analysis for confirmation screens

Analysing confirmation data

Standard mass screens were run using a single well for each compound, dosed at 10 μM in the assay. Confirmation screens were designed to identify whether compounds labelled as a *potential hit* in the standard screen are truly active; a range of concentrations was used for each compound, and replicates of each experiment were run. The list of confirmation screens is given in Appendix A.6.

Multiple replicates and concentrations allow an activity/concentration curve to be built for each compound/assay combination; these were inspected visually for evidence of preferential activity against the parasite strain rather than the human control. The curves were examined to determine which compounds were clearly active using the confirmation screen data; these observations could then be used in combination with the discrete curve data for each well to produce further rules to classify their relative activity.

Two main sets of experimental conditions have been used: earlier experiments used 8 replicates across a concentration range of 1 to 20 μM; this was later changed to 4 replicates of concentration range 0.5 to 10 μM. The change in conditions allows more compounds to be run on each plate, with the added benefit of reducing the amount of chemical compound required for each confirmation curve; the reduction in concentration range was also designed to be more representative of a practical therapeutic dose. Changing the conditions will have an effect on the concentration curve: a reduced number of replicates will mean more noise, and a reduced concentration range will lower the likelihood of a compound appearing active (although it may realise extended activity at lower concentrations).

On considering these variations, it was decided to build a separate set of rules for each population. The data from TS6 (CS_77_7_16_20110529120129) was used for the earlier approach, and the data from TS7 (CS_80_3_22_20110714113111) for the latter. Subsequent results were then categorised according to the rules for an individual experiment, and these were analysed using Weka (J48 decision tree, 10 fold cross validation) in combination with the visual results for each compound's concentration curve. The concentration curves were categorised visually into ***active***

versus the sapphire fluorophore, ***active versus the venus fluorophore***, ***weak activity***, ***toxic***, and ***inactive***. It was possible to grade compounds with additional details, such as being active against two targets, or being a discrete hit at low concentration but toxic at high concentration.

The following rules were observed when splitting visual observations into simpler parcels such as hit (active/weak), toxic, and inactive:

Category	**Rule**	**General interpretation/ comments**
TS6 Venus hit	v_hit $\leq$ 3 [23 inactive/4 weak/2 active] v_hit > 3 [23 active/11 inactive] For active compounds: 46/63 = 73% prediction success 2/25 = 8% false negatives	If more than three of the 40 individual venus curves are hits, then the compound is active against the venus target.
TS6 Sapphire hit	s_hit $\leq$ 3 [4 inactive/1 active] 3 < s_hit $\leq$ 5 [2 weak] s_hit > 5 [active 53/inactive 3] For active compounds: 58/63 = 90% prediction success 1/54 = 2% false negatives	If more than five of the 40 individual sapphire curves are hits, then the compound is active against the sapphire target.
TS6 Toxic	c_hit $\leq$ 4 [48 inactive/2 active] c_hit > 4 [11 active/2 inactive] For active compounds: 59/63 = 92% prediction success 2/11 = 18% false negatives	If more than four of the 40 individual cherry curves are hits, then the compound is active against the cherry target, and is classed as possibly toxic.

Table 3.12: Decision tree rules for TS6 confirmation results

The data set for TS6 consisted of 63 compounds, which were initially chosen for the confirmation step due to potential activity versus the PvDHFR target (sapphire fluorophore). The decision tree rules for the sapphire hits were therefore of most interest for this screen: they describe sapphire activity very well, with only one visually active classification not being selected out of 54 so classified.

The rules developed for the TS6 venus target were less clear, relating to the borderline nature of the confirmation curves. A relatively high number of false positives were selected by the rules (visually classified as inactive); a visual re-examination of the curves suggested that activity versus the venus target may have been missed due to the relatively strong signal for the sapphire target in the same well. This shows the strength of taking an objective machine learning approach for building rules for complex data analysis.

Category	**Rule**	**General interpretation/ comments**
TS7 Venus hit	v_hit $\leq$ 3 [17 inactive/2 active] v_hit > 3 [26 active/8 inactive] For active compounds: 43/53 = 81% prediction success 2/26 = 8% false negatives	If more than three of the 20 individual venus curves are hits, then the compound is active against the venus target.
TS7 Sapphire hit	Insufficient hits for J48 decision tree to work	
TS7 Toxic	c_hit $\leq$ 3 [39 inactive/1 active] c_hit > 3 [9 active/4 inactive] For active compounds: 48/53 = 91% prediction success 1/9 = 11% false negatives	If more than three of the 20 individual cherry curves are hits, then the compound is active against the cherry target, and is classed as possibly toxic.

Table 3.13: Decision tree rules for TS7 confirmation results

The 53 compounds used to build the TS7 decision tree rules were predominantly selected for their potential activity versus PvRdhfr (venus flurophore). In general, the strength of the confirmation curves was weaker than with the earlier TS6 set, hence the higher proportion of false positives (8 out of 34 compounds labelled as active versus PvRdhfr).

3.2.6 Labelling activity of compounds in confirmation screens

Applying the above rules, the simplest approach was to classify a compound as active and/or possibly toxic based on the cumulated screen rules score; under the mass screen rules an active compound is [score = 2] for a full hit, or [score = 1] for activity based on lagtime or doubletime. Therefore, [score > 9] was taken as the limit above which a compound is defined active or toxic for TS6-type populations, and [score > 7] for TS7-type populations. These limits were programmed in R, and this code was then extended to provide ranked lists of active compounds for each target strain, split into "No toxicity indicated" and "Possibly toxic" sections.

Population type	Hs fluorophore score	Parasite flurophore score	Toxicity	Parasite activity
8 replicates, 1 – 20 μm	≤ 9	≤ 9	No toxicity indicated	Inactive
		> 9		Active
	> 9	≤ 9	Possibly toxic	Inactive
		> 9		Active
4 replicates, 0.5 – 10 μm	≤ 7	≤ 7	No toxicity indicated	Inactive
		> 7		Active
	> 7	≤ 7	Possibly toxic	Inactive
		> 7		Active

Table 3.14: Generalised decision tree rules for confirmation results

3.2.7 Verification of labelling rules for confirmation screens

Several confirmation screens were assessed both visually and by using the rules built with the data from TS6 (CS_77_7_xx) and TS7 (CS_80_3_xx), then generalised in Table 3.14. The results in Table 3.15 show how many compounds were visually classified as active or inactive against the parasite targets, and whether there was an indication of general toxicity. The columns showing the classification under the decision tree rules show the number of items classified and mis-classified against the 'inactive' and 'active' labels respectively.

In general, the rules work well when comparing such classified activity against the visual results for each confirmation screen. Some difference is seen concerning borderline toxicity; it is likely that this could be improved by adding an additional rule to the ranking mechanism for active compounds, thereby allowing strongly active compounds with borderline toxicity to be more visible. It might also be possible to improve the rules by including either lagtime or doubling time.

Screen	Classification	Visual		Decision tree rules		Comments
		Inactive	Active	Inactive	Active	
TS3_63_2	Sapphire	8	50	8/1	50/1	Rules pick up more borderline toxic compounds in TS3; this accounts for additional venus hits too. Main activity is reproduced very favourably.
	Venus	52	6	46/0	12/6	
	Toxic	54	4	45/1	13/10	
TS4_64_4	Sapphire	20	42	23/5	39/2	Rules suggest more borderline active & toxic compounds. Good agreement with main activity.
	Venus	47	15	33/2	29/16	
	Toxic	51	11	35/1	27/17	
TS5_71_2	Cherry	24	36	18/2	42/8	Rules suggest more borderline active & toxic compounds. Good agreement with main activity.
	Venus	22	38	15/2	45/9	
	Toxic	37	23	33/4	27/8	
TS6_cherrypick	Sapphire	15	33	20/6	28/1	All four TS6 screens showed good general agreement between visual and rule-based analyses; the disagreements were with weak visual actives and with borderline toxic results using the rules.
	Venus	32	16	30/2	18/4	
	Toxic	42	6	36/0	12/6	
TS6_77_3	Sapphire	85	11	88/3	8/0	
	Venus	91	5	90/1	6/2	
	Toxic	91	5	89/1	7/3	
TS6_77_4	Sapphire	80	22	90/10	12/0	
	Venus	90	12	96/6	6/0	
	Toxic	98	4	96/1	6/2	

Table 3.15 (part 1): Confirmation screen data – visual and rule-based activity classification

Screen	Classification	Visual		Decision tree rules		Comments
		Inactive	Active	Inactive	Active	
TS6_77_5	Sapphire	72	14	84/12	2/0	
	Venus	81	5	82/2	4/1	
	Toxic	80	6	81/2	5/1	
TS6_77_7	Sapphire	5	58	6/2	57/1	Rules pick up more borderline toxics, & sees other venus that might have been swamped by strong visual sap signal
	Venus	34	29	23/1	40/12	
	Toxic	56	7	44/0	19/12	
TS7_80_4	Sapphire	36	17	40/5	13/1	Good agreement throughout. Extra visual venus hits should really just be labelled as toxic.
	Venus	10	43	19/9	34/0	
	Toxic	43	10	40/1	13/4	
PGK1_72_2	Sapphire	23	36	20/2	39/5	General agreement good. Lots of well labelled toxics too.
	Venus	22	37	17/2	42/7	
	Toxic	33	26	30/3	29/6	
PGK2_74_2	Sapphire	37	25	45/11	17/3	Good agreement. Main disagreement around borderline/weak activity.
	Venus	30	33	35/8	27/2	
	Toxic	50	13	53/5	9/1	
NMT1_78_2	Sapphire	16	45	22/8	39/2	Good agreement. Main disagreement around weak activity.
	Venus	27	34	32/6	29/1	
	Toxic	59	2	60/1	1/0	
NMT2_79_2	Sapphire	36	25	43/9	18/2	Good agreement. Main disagreement around weak activity.
	Venus	19	42	24/7	37/2	
	Toxic	47	14	50/3	11/0	

Table 3.15 (part 2): Confirmation screen data – visual and rule-based activity classification

3.3 *In vivo* experiments with yeast-hosted targets

Robot Scientist Eve ran mass screen experiments on triple strain sets between December 2010 and January 2012. The results from these experiments provided leads for subsequent confirmation/cherry-pick screening, which duly ran between December 2010 and June 2012, at which point Eve was relocated to Manchester.

3.3.1 Screens and cherrypicking

Data from Eve

Eve produces a comma-separated-variable data file (.csv) for each plate:

384 wells per plate: each plate is 16 rows (labelled A to P) × 24 numbered columns.

This typically represents 320 compounds and 64 control wells. The control wells (up to 20 positive controls per plate; the remainder are negative controls) are typically located in the outer two columns for each plate (1, 2, 23 & 24) to help to counter edge-drying effects.

A line of data is recorded for each well, with 25 parameters per well (29 for confirmation screens). Assuming no empty wells have occurred (for which the data is removed at source), Eve therefore produces a 384×25 dataframe for each mass screen plate.

Eve saves the .csv file in the format:

MS_77_1_22_20110316174043.csv

MS = mass screen (CS = cherry/confirmation screen)

77 = code number for the assay in use

1 = code for the iteration number of this assay

22 = rolling number of experiments conducted on Eve to date

20xx = date/time stamp (2011, 16 March, 17:40:43)

The lists of mass and confirmation screens are given in Appendix A.2 and A.6 respectively.

3.3.2 Mass screen results

The decision tree rules derived for mass screens were subsequently applied for all such experiments run by Eve in Aberystwyth. A total of 17 parasite targets (9×DHFR, 4×NMT and 4×PGK; see Appendix A.1 for details) have been evaluated during the course of 12 mass screens (Appendix A.2). With the exception of TS6 which was split over two screens, all mass screens contained all of the compounds available in the Maybridge Hitfinder Library and the JHCCL.

The last two mass screens run in Aberystwyth (TS8 and TS9) both contained the SaDHFR target on the sapphire and venus fluorophores; for both screens, problems were experienced with fluorescence measurements. An attempt was made to cherry-pick compounds using the data from TS9, but there was too much noise to provide usable confirmation results.

Appendix A.3 shows the number of active compounds versus each target. Each active compound is categorised as either a clean hit (i.e. no activity versus the two other targets in the screen), a co-hit with the other parasite target, or as *possibly toxic* (a hit, but also with a weak signal suggesting activity versus the Hs strain).

Example – activity versus PvDHFR in TS6

203 compounds were identified that had discrete activity versus PvDHFR in the TS6 mass screen, with 81 that were jointly active versus PfDHFR and a further 19 considered active but possibly toxic.

The ideal behaviour for a strong candidate would be a large reduction in the growth of the target strain, with a corresponding overgrowth of the Hs strain. This behaviour could be indicated by measuring the relative growth ratio at the mass screen stage, as shown in Table 3.16. The decision tree scores in this table can be used to indicate whether activity is against a single target, or if there is potential for a multiple target effect or toxicity.

For each mass screen, the candidates offering the best target activity potential were ranked for further evaluation in a confirmation screen (section 3.3.3). This ranking was conducted by taking their activity relative to the other endogenous targets.

Eve ID	Growth ratio, DHFR				Decision tree score				$\frac{Pv\ growth}{Total\ growth}$	$\frac{Pv\ growth}{Hs\ growth}$
	Hs	Pv	Pf	Total	Hs	Pv	Pf	Total		
978	3.17	0.12	0.15	3.44	0	2	2	4	0.04	0.04
9082	0.96	0.16	0.77	1.89	1	2	2	5	0.08	0.16
9499	1.28	0.26	0.89	2.43	0	2	0	2	0.11	0.21
10838	0.94	0.20	0.77	1.90	1	2	2	5	0.10	0.21
8766	1.21	0.27	0.80	2.29	0	2	0	2	0.12	0.22
10371	1.20	0.29	0.66	2.15	1	2	2	5	0.13	0.24
7091	1.76	0.44	0.68	2.87	1	2	2	5	0.15	0.25
15497	1.89	0.51	0.49	2.89	1	2	2	5	0.18	0.27
3466	1.29	0.37	0.72	2.38	0	2	2	4	0.15	0.28
5829	1.27	0.36	0.90	2.54	0	2	0	2	0.14	0.29
7352	0.40	0.12	2.19	2.71	2	2	1	5	0.04	0.29
12803	1.95	0.58	0.74	3.26	0	2	2	4	0.18	0.30
16914	1.63	0.53	2.29	4.44	2	2	1	5	0.12	0.32
13015	1.99	0.64	2.11	4.74	0	2	1	3	0.14	0.32
17167	1.12	0.36	0.70	2.18	1	2	2	5	0.17	0.33
14244	0.80	0.26	2.74	3.81	1	2	1	4	0.07	0.33
3978	0.96	0.32	2.02	3.30	2	2	1	5	0.10	0.33
12054	1.27	0.44	0.95	2.67	0	2	0	2	0.17	0.35
12913	1.38	0.48	3.74	5.61	1	2	1	4	0.09	0.35
5833	1.82	0.64	1.21	3.67	0	2	1	3	0.17	0.35

Table 3.16: Top 20 PvDHFR active candidates from TS6 mass screen, by relative growth

These 20 candidates all showed growth less than 35% that of the HsDHFR target during the course of the experiment, with most also displaying overgrowth of the latter. Note: Eve ID 978 is Pyrimethamine, the main positive control for the unmutated plasmodium species.

The low growth for the Hs target for one candidate (Eve ID 7352) suggests that this is likely to have general toxicity, and two other candidates also had both the lagtime and doubling time indicators of weak toxicity, but reasonable Hs growth (Eve ID 3978 & 16914).

Mass screen negative control statistics

The negative control statistics are displayed graphically in Appendix A.5. The comments raised for TS3 (see section 3.2.1, where noisy data was identified due to incorrect make up of the growth media) can be related to the kick in the plot of standard deviation versus plate number. There is a distinct rise in standard deviation (SD) for [sapphire=blue] and [venus=green].

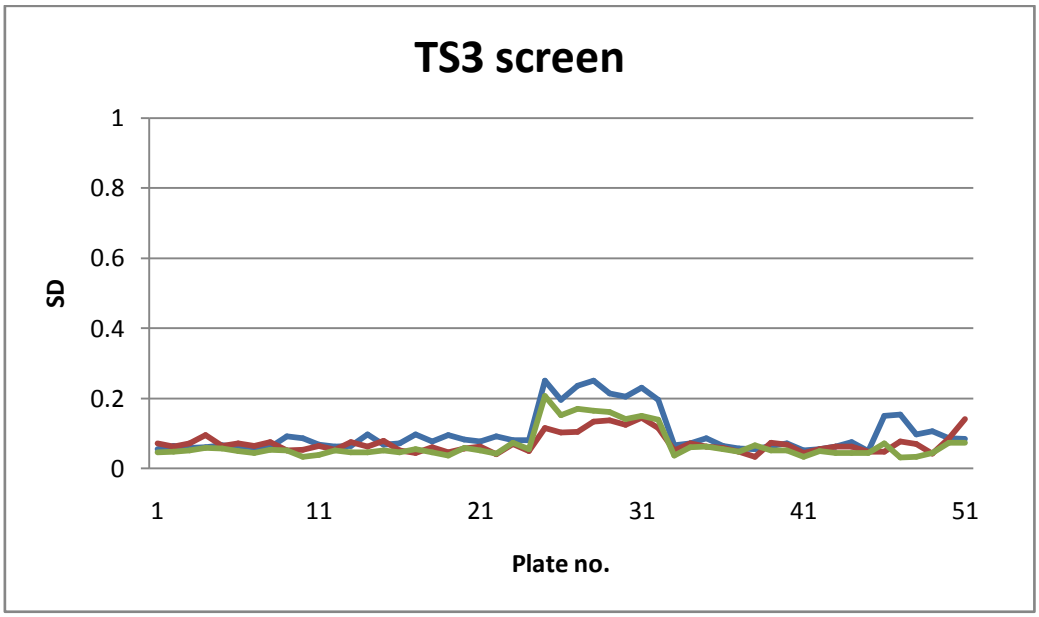


Figure 3.6: Standard deviation of TS3 negative controls, by plate

This noise was high enough to cloud data interpretation for these plates; plates in other screens also showed high levels of noise, and in some cases a large proportion of the full screen was affected (e.g. TS8, TS9, PGK-2). Unfortunately the causes could not be investigated for the latter mass screens, as the level of technical

support in Aberystwyth had reduced significantly by the time they were run. The screens with high noise coincided with targets that returned lower proportions of hit compounds; if reasons for high noise could be isolated and countered, this might allow more screen data to be available, although it is highly likely that the affected screens would need to be re-run.

3.3.3 Confirmation screen results

A full list of the confirmation screens is given in Appendix A.6.

Appendix A.8 gives a full list of Maybridge Hitfinder compounds with activity against the relevant parasite strains during confirmation screens. Appendix A.9 records equivalent activity lists for all active JHCCL compounds.

The example compounds from Table 3.16 that potentially indicated activity versus the PvDHFR strain have been followed through the confirmation screen process:

17 of the 20 strong PvDHFR candidates were run as part of one of the TS6 confirmation screens (CS_77_7_16_...); the three compounds not in this screen were 12913, 17167 and 978. In accordance with the confirmation screen rules, the compound activity scores and classifications are as follows:

Eve ID	Target score			Active vs target		Toxicity
	Hs	Pv	Pf	Pv	Pf	
9082	2	80	14	Yes	Yes	None indicated
9499	8	80	10	Yes	Yes	None indicated
10838	21	80	31	Yes	Yes	Possibly toxic
8766	0	80	2	Yes		None indicated
10371	0	80	76	Yes	Yes	None indicated
7091	0	22	0	Yes		None indicated
15497	8	20	21	Yes	Yes	None indicated
3466	16	36	18	Yes	Yes	Possibly toxic
5829	0	72	0	Yes		None indicated
7352	10	13	8	Yes		Possibly toxic
12803	16	36	42	Yes	Yes	Possibly toxic
16914	33	38	26	Yes	Yes	Possibly toxic
13015	9	28	8	Yes		None indicated
14244	14	48	19	Yes	Yes	Possibly toxic
3978	14	56	15	Yes	Yes	Possibly toxic
12054	7	64	9	Yes		None indicated
5833	30	36	21	Yes	Yes	Possibly toxic

Table 3.17: Confirmation screen of 17 strong TS6 PvDHFR active candidates

The confirmation curve for each of these compounds is given in Appendix A.7.

Example curves for stronger and weaker candidates, and for those exhibiting possible toxicity are shown in Figures 3.7 to 3.9; in these figures, HsDHFR is red, PvDHFR is blue, and PfDHFR is green.

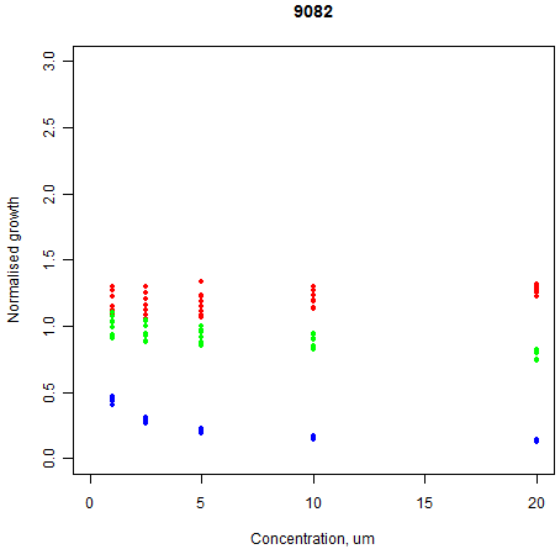


Figure 3.7: Confirmation curves for Eve ID 9082

The confirmation curves for 9082 shows strong growth inhibition for the PvDHFR strain, with a weaker signal for PfDHFR.

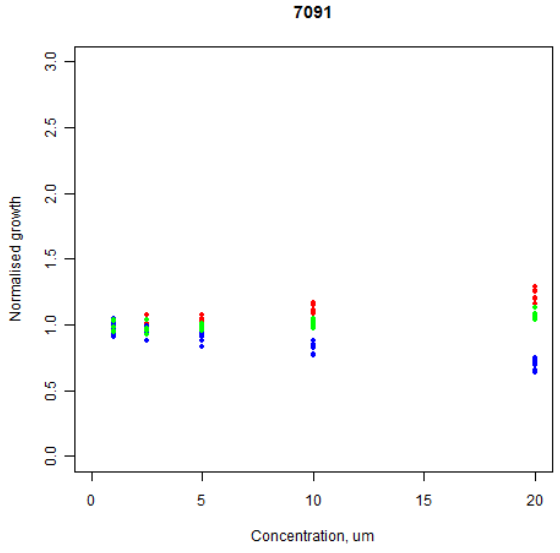


Figure 3.8: Confirmation curves for Eve ID 7091

The PvDHFR concentration curves for 7091 shows weak growth inhibition, with little effect on the other two strains.

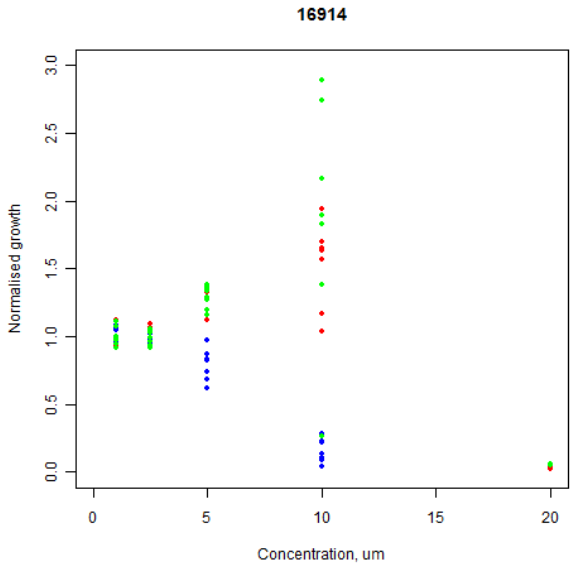


Figure 3.9: Confirmation curves for Eve ID 16914

All three curves for 16914 are very noisy, which is a good indicator that the compound is generally toxic versus these targets.

Confirmation screen negative control statistics

The negative control statistics are displayed graphically in Appendix A.10. A higher variability of in-plate standard deviation was seen than during the mass screens.

As more progress is made with the development of Eve's data handling and analysis software, these negative control data can be used to define limits at which to reject results, or at least to label them as in need of further verification. More repeat runs of mass and confirmation screens would be needed to construct such limits.

3.3.4 Testing an expansion seeded from confirmed hits

After the first TS6 confirmation screen, it was noted that two adjacent compounds in the MaybridgeHF library (no.9081 & 9082) were active versus PvDHFR and had structural similarities. A simple structural fingerprint search revealed other similar library compounds.

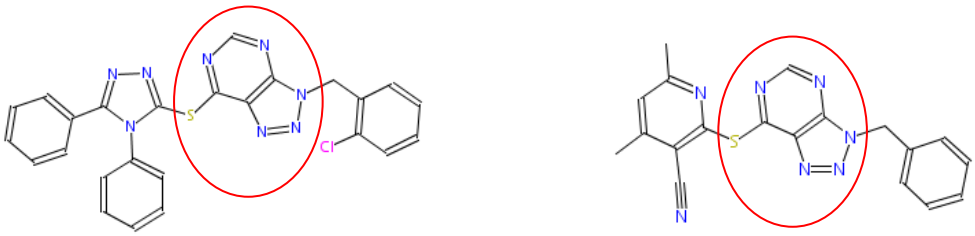


Figure 3.10: Core structural similarities for 9081 & 9082

The levels of activity in the mass and confirmation screens (where tested) of the three additional compounds are given in the Table 3.18. Only one of the five MaybridgeHF compounds was categorised as inactive after the mass screen.

Subsequently, a search of the full Maybridge library (available as an online resource) was conducted, and 7 other candidates were found that had strong Tanimoto Similarity to 9081/HTS12148. Six of these candidates were purchased for confirmation screen testing, and the other was not available; their confirmation growth curves are displayed in Table 3.19.

Two candidates (18013 & 18015) were inactive in the confirmation screen; interestingly, they were smaller molecules with no large functional group attached to the thiol bridge, and could be described as small molecules for drug lead development in this context. The other four compounds were active versus PvDHFR, although 18014 could be better described as active/possibly toxic.

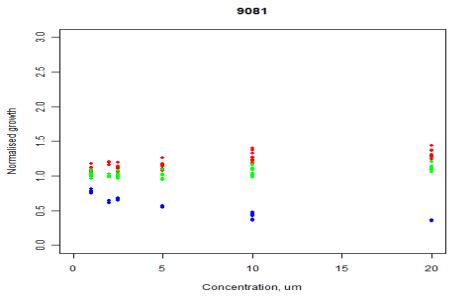
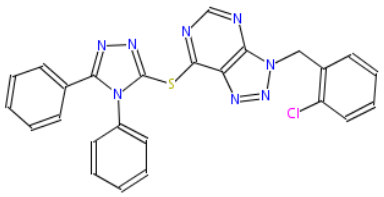
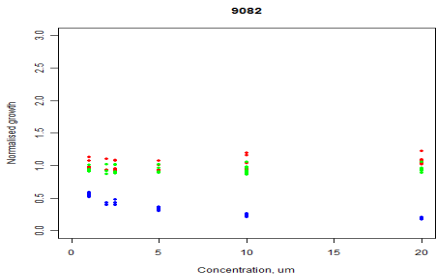
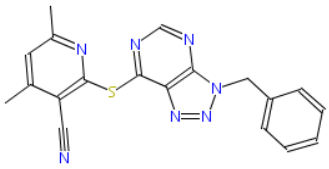
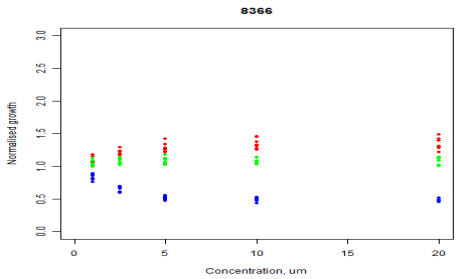
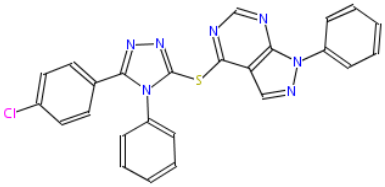
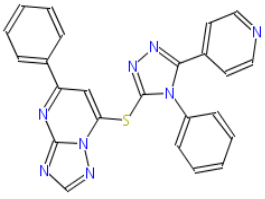
Eve ID Maybridge ID Tanimoto Activity	Confirmation growth curves	Molecular structure/SMILES
9081 HTS12148 - Confirmed PvDHFR hit		 N1(N=Nc2c1ncnc2SC1=NN=C(c2ccccc2)N1c1ccccc1)Cc1ccccc1Cl
9082 hfHTS12152 0.580247 Confirmed PvDHFR hit		 N1=NN(c2c1c(ncn2)Sc1nc(cc(c1C#N)C)C)Cc1ccccc1
8366 hfHTS 07614 0.747664 Confirmed PvDHFR hit		 N1(c2ccccc2)C(=NN=C1c1ccc(cc1)Cl)Sc1c2c(ncn1)N(c1ccccc1)N=C2
9046 hfHTS 11966 0.62931 TS3 Pv hit	Unknown (not tested)	 N1=CN=C2N1C(=CC(=N2)c1ccccc1)SC1=NN=C(c2ccncc2)N1c1ccccc1

Table 3.18: Activity of Maybridge HF compounds similar to Eve ID 9081/9082

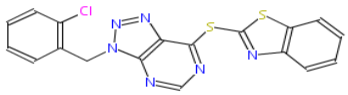
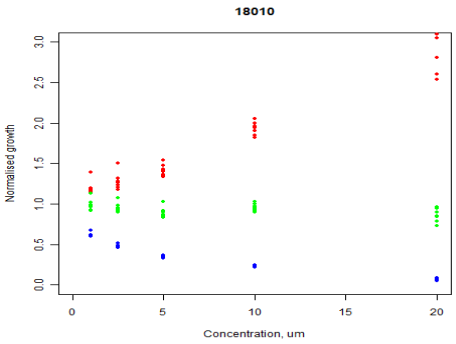
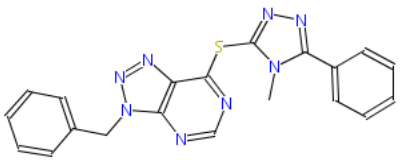
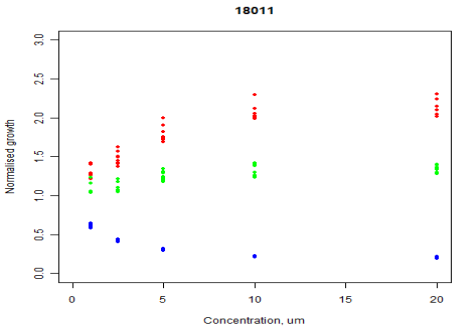
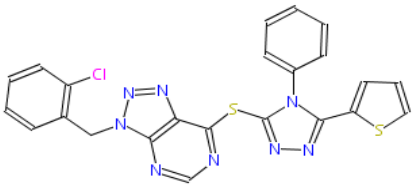
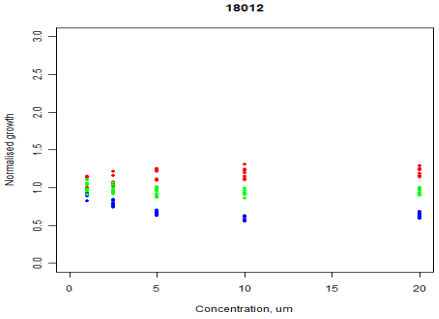
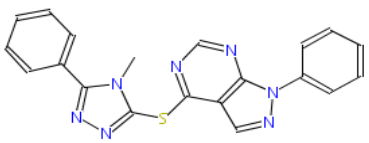
Eve ID Maybridge ID Tanimoto vs 9081 Activity	Confirmation growth curves	Molecular structure/SMILES
9080 hfHTS 12146 0.619247 TS3 inactive	Unknown (not tested)	 N1(c2c(c(ncn2)SC2=Nc3c(cccc3)S2)N=N1)Cc1ccccc1Cl
18010 HTS12151SC 0.918367 Confirmed PvDHFR hit		 Cn1c(Sc2ncnc3n(Cc4ccccc4)nnc23)nnc1c5ccccc5
18011 HTS12147SC 0.841202 Confirmed PvDHFR hit		 Clc1ccccc1Cn2nnc3c(Sc4nnc(c5cccs5)n4c6ccccc6)ncnc23
18012 HTS07613SC 0.680751 Confirmed PvDHFR hit		 Cn1c(Sc2ncnc3n(ncc23)c4ccccc4)nnc1c5ccccc5

Table 3.19: Activity of Maybridge compounds similar to 9081/9082

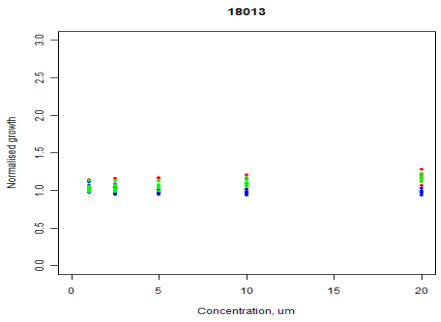
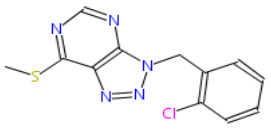
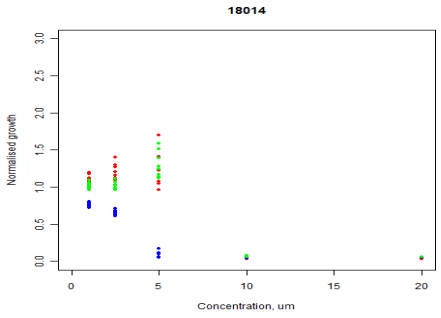
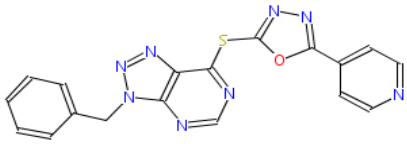
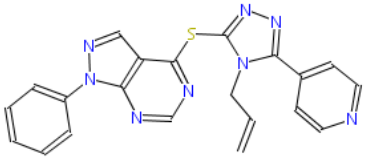
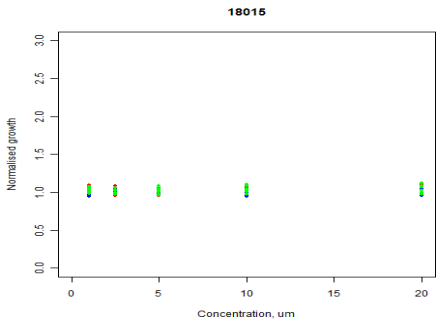
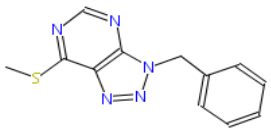
Eve ID Maybridge ID Tanimoto vs 9081 Activity	Confirmation growth curves	Molecular structure/SMILES
18013 HTS12149SC 0.663265 Weak PvDHFR activity		 CSc1ncnc2n(Cc3ccccc3Cl)nnc12
18014 HTS12150SC 0.644068 Pv hit/toxic		 C(c1ccccc1)n2nnc3c(Sc4nnc(o4)c5ccnc c5)ncnc23
Not available HTS09074SC 0.625551 Unknown	Unknown (not available)	
18015 HTS12154SC 0.612245 Inactive		 CSc1ncnc2n(Cc3ccccc3)nnc12

Table 3.20: Activity of Maybridge full library compounds similar to 9081/9082

3.3.5 Confirmed activity for JHCCL compounds

In addition to testing JHCCL drugs in confirmation screens (see Appendix A.9) against a single parasite-protein combination, the results were scrutinised to understand if any background effects might be suggesting false positives.

It had been noted earlier that some MaybridgeHF compounds were repeatedly identified as active versus a particular fluorophore in otherwise unrelated screens; this suggested that fluorophore-specific effects should be evaluated once potential hits had been identified.

The intrinsic value of finding a hit amongst the pre-approved JHCCL drugs is high. Many of these potential hit compounds were subsequently tested across a wide range of assays; this allowed discrimination of the compound activity versus unrelated proteins, thereby determining whether the effect might be protein-specific or an indicator of more generalised toxicity.

A total of 13 JHCCL drugs were identified with activity versus at least one parasite target; Methotrexate was added to this list (another approved drug, present as a positive control). The presence of each drug in multiple confirmation screens allowed statistical analysis of background effects to be considered.

General performance versus either the fluorophore or the protein was evaluated using two-sample *t* tests and Wilcoxon Ranked Sign Tests using all the concentration curve data at 5 µM and 10 µM. For identifying whether a compound was active versus a parasite-protein target independently from the fluorophore, the active data was compared to all data from inactive results. Where more than one active target was identified, these results were separated from the inactive data, and tested for their own specific activity.

As an example, TNP-470 was labelled as active versus PvDHFR in a confirmation screen, but not labelled as active versus any other targets tested (Hs/PfR/Tc& PfDHFR, Hs/Tb/Pv/Sm/TcNMT). Testing the PvDHFR results versus all other (inactive) DHFR results showed the activity to be Pv-specific; similar statistical analysis of the results versus all TNP-470 experiments run against the sapphire fluorophore again showed independent activity (see Table 3.23).

Drug name	JHCCL ID, SM_JHU_xx	Independence of confirmed activity
Azaribine	904	Active versus **HsDHFR**, **TcDHFR** & **PfDHFR**, independent of fluorophore. [Note: all HsDHFR compared to all HsNMT]
Triclosan	10450	PvDHFR, TbNMT & SmDHFR activity is not independent of fluorophores.
Bithionoloxide	1305	**TbNMT** activity is independent of fluorophore. PvDHFR activity is not independent of fluorophore.
Satraplatin	9003	**PvDHFR** & **TcNMT** activity is independent of fluorophore.
Apomorphine	715	**PvDHFR** activity is independent of fluorophore.
Chloroquinaldol	2095	PvDHFR activity is not independent of fluorophore.
Lamotrigine	5529	PvDHFR activity is borderline independent of fluorophore.
Closantel	2430	PvNMT, TbNMT: insufficient data to test independence
Dichlorophen	3038	**TbNMT** activity is independent of fluorophore.
Methotrexate	-	**TcDHFR** activity is independent of fluorophore.
Demacarium Bromide	2831	**SmNMT** activity is independent of fluorophore.
Meclocycline	6047	**SmNMT** activity is independent of fluorophore.
Stanozolol	9409	HsDHFR: insufficient data to test for independence
TNP-470	10251	**PvDHFR** activity is independent of fluorophore.

Table 3.21: Generalised independence results for active JHCCL compounds

Compound	Enzyme	Strain	Fluorophore	Versus inactive enzyme analogues						Versus inactive fluorophore analogues					
				10um			5um			10um			5um		
				t-stat	df	WRST-p	t-stat	df	WRST-p	t-stat	df	WRST-p	t-stat	df	WRST-p
Azaribine	DHFR	Hs	mCherry	-4.97	78	1.13×10^{-4}	-4.27	78	8.77×10^{-4}	-5.40	42	2.57×10^{-6}	-2.28	42	0.0486
	DHFR	Hs	Sapphire												
	DHFR	Tc	Venus	-13.5	62	2.87×10^{-14}	-12.2	62	4.96×10^{-12}	-11.4	38	4.60×10^{-7}	-7.28	38	8.59×10^{-8}
	DHFR	Pf	Venus	-9.71	54	2.82×10^{-9}	-9.78	54	1.69×10^{-8}	-8.45	30	3.80×10^{-7}	-6.29	30	3.61×10^{-6}
Triclosan	DHFR	Pv	Sapphire	-5.94	102	1.95×10^{-7}	-8.08	102	2.57×10^{-10}	0.513	38	0.859	-1.43	38	0.1489
	DHFR	Pf	Venus	-1.28	94	0.0996	-2.96	94	5.91×10^{-4}	-1.69	50	0.0526	-3.87	50	1.90×10^{-4}
	NMT	Tb	Sapphire	-2.90	26	3.71×10^{-3}	-3.79	26	1.17×10^{-3}	-0.339	18	3.71×10^{-3}	-0.566	18	0.494
	NMT	Sm	Sapphire	-2.23	26	2.89×10^{-3}	-3.43	26	8.56×10^{-6}	0.407	18	0.834	0.372	18	0.452
Bithionoloxide	DHFR	Pv	Sapphire	-1.21	98	0.406	-3.93	98	5.55×10^{-5}	-0.142	42	0.920	0.698	42	9.09×10^{-3}
										2.26	50		0.833	50	
	DHFR	Pf	Venus	-1.17	90	0.174	-3.50	90	5.92×10^{-4}	3.02	46	0.0130	4.09	42	3.76×10^{-3}
	DHFR	PfR	mCherry	-4.00	94	7.08×10^{-3}	-2.52	94	0.153	-2.17	58	0.0171	-1.12	58	0.0875
	DHFR	PfR	Venus							2.66	42	0.0171	0.627	46	0.573
	NMT	Tb	Sapphire	-4.70	26	4.31×10^{-5}	-4.08	26	2.98×10^{-4}	-3.64	42	5.25×10^{-5}	-4.06	42	6.61×10^{-8}
	NMT	Pv	Venus	-3.84	26	5.12×10^{-4}	-2.76	26	0.0183	-1.21	38	0.0834	-2.28	38	0.0515
Satraplatin	DHFR	Pv	Sapphire	-11.4	38	3.18×10^{-11}	-10.7	38	8.87×10^{-13}	-3.37	14	1.66×10^{-3}	-3.26	14	1.82×10^{-4}
	DHFR									-2.97	14		-3.19	14	
	DHFR	PfR	Venus	-4.40	30	3.18×10^{-11}	-3.92	30	3.17×10^{-4}	-2.62	18	2.16×10^{-3}	-2.07	18	0.452
	NMT	Tc	Venus	-3.30	22	1.88×10^{-4}	-3.38	22	1.88×10^{-4}	-5.53	14	1.10×10^{-3}	-3.39	14	1.110×10^{-3}
Apomorphine	DHFR	Pv	Sapphire	-6.66	62	1.18×10^{-5}	-5.78	62	1.17×10^{-4}	-3.32	26	8.34×10^{-3}	-3.12	26	9.71×10^{-3}
Chloroquinaldol	DHFR	Pv	Sapphire	-3.92	82	4.45×10^{-6}	-3.83	82	2.06×10^{-6}	0.487	34	0.934	-0.365	34	0.753

Table 3.22: Statistical analysis of independent protein/fluorophore activity for active JHCCL compounds (part I)

Compound	Enzyme	Strain	Fluorophore	Versus inactive enzyme analogues						Versus inactive fluorophore analogues					
				10um			5um			10um			5um		
				t-stat	df	WRST-p	t-stat	df	WRST-p	t-stat	df	WRST-p	t-stat	df	WRST-p
Lamotrigine	DHFR	Pv	Sapphire	-5.76	58	1.73×10^{-4}	-4.72	58	7.20×10^{-3}	-2.22	26	0.0120	-1.98	26	0.0387
2,4,6-tribromophenol	DHFR	Pv	Sapphire	-5.13	30	1.73×10^{-4}	-5.45	30	7.20×10^{-3}	2.21	14	0.0121	1.60	14	0.0387
	NMT	Tb	Sapphire	-5.60	19	1.50×10^{-3}	-4.23	19	1.88×10^{-4}	-7.97	9	2.06×10^{-3}	-3.65	9	4.13×10^{-4}
Closantel	NMT	Pv	Venus	Insufficient data to test for cross-strain/fluorophore independence											
	NMT	Tb	Sapphire												
Dichlorophen	DHFR	Tc	Venus	-0.971	58	0.254	-0.680	58	0.646	-2.01	26	0.0421	-1.27	26	0.291
	NMT	Tb	Sapphire	-3.53	22	1.88×10^{-4}	-4.68	22	1.88×10^{-4}	-2.60	26	3.71×10^{-4}	-4.30	26	6.84×10^{-4}
Methotrexate	DHFR	Tc	Venus	-3.23	235	0.0644	-3.07	238	0.0749	-3.57	92	0.0918	-3.46	92	0.0461
Demacarium Bromide	NMT	Sm	Sapphire	-5.17	46	5.43×10^{-6}	-6.57	46	3.23×10^{-6}	-5.33	26	1.93×10^{-5}	-6.05	26	1.93×10^{-5}
Meclocycline	NMT	Sm	Sapphire	-3.58	46	5.12×10^{-4}	-3.52	46	6.69×10^{-4}	-3.58	26	0.0134	-3.58	26	0.0134
Stanozolol	DHFR	Hs	mCherry	Insufficient data to test for cross-strain/fluorophore independence											
TNP-470	DHFR	Pv	Sapphire	-4.84	94	9.05×10^{-3}	-3.95	94	5.17×10^{-3}	-3.21	38	8.70×10^{-3}	-3.07	38	4.60×10^{-4}

Notes

t-stat: Two-sample t test statistic

df: Degrees of freedom

WRST-p: Wilcoxon Ranked Sign Test, *p*-value

Red text: t statistic < -2.00 ($p < 0.025$)

Table 3.23: Statistical analysis of independent protein/fluorophore activity for active JHCCL compounds (part II)

3.4 A comparison of rule development methods for Eve's data

3.4.1 Summary

The predictions of Eve's decision tree-based rules have been compared to those made by an alternative method provided by the School of Biological Sciences, University of Cambridge. Both approaches are designed to predict potential hit compounds following a mass screen, and to eliminate those compounds likely to be toxic or autofluorescent; analysis has shown that they have limited overlap.

Eve's data analysis method demonstrates better discrimination between active and toxic compounds for individual screens, and also eliminates autofluorescent compounds that are otherwise missed. By incorporating measurements of lagtime and doubling time, Eve's rules demonstrate a distinct advantage in their ability to identify compounds otherwise labelled as false negatives by the Cambridge method.

The method provided by Cambridge uses data from across all screens during the selection process; this helps to eliminate compounds which are generally toxic or have repeatedly caused processing problems, and provides useful ideas for post-evaluation of data from multiple screens. However, by only using end-of-test growth ratios, it misses information gleaned by Eve's rules that use growth rates and lagtimes, and consequently its ability to identify screen-specific activities is restricted.

Ideally, a combined method could be constructed that incorporates the cross-screen data analysis of the Cambridge method with the additional information inputs used by the original decision tree rules. This might itself be an Active Learning exercise for Eve to maintain.

3.4.2 Introduction

Two separate approaches have been used to define hits using Robot Eve's mass screen results, classifying each compound well in terms of relative activity and effects from noise. The Cambridge method exploited the benefits of using all data across all available screens; this enabled it to build filters that highlighted compound wells that exhibited repeated problematic patterns e.g. autofluorescent & fluorophore-specific effects.

In contrast, the rules originally derived for Eve were built using supervised decision tree learning based on visual interpretation of only the original triple strain mass screen (as described in section 3.2).

Cambridge's filters (based on analysis of end-of-test growth):

- Toxic = {max growth < 0.70 for all three strains across all screens}
- Autofluorescent = {mcherry, sapphire or venus > 1.2 in all screens}
- Fluorophore specific = {growth < 0.60 for a specific fluorophore in all screens}
- Problem well = {average value in all channels in a specific screen < 50% of average value in all other screens}
- Hit = {parasite strain growth / human strain growth < 0.5}

Eve's filters (see section 3.2.4):

- Autofluorescent = {initial fluorescence is more than 8% through the growth range for the negative control}
- Potential hit = {growth < 80% of the negative control}
- Possibly active = {growth > 80% of the negative control but
  (a) lagtime > 4 hours after the negative control or
  (b) doubling time > 50% above the negative control}

Eve's labelling process:

1. Mark and remove the fluorescent compounds.
2. If compounds are potential hits on all channels (mcherry, sapphire, venus), mark them as ***definitely toxic***.
3. Scoring a ***potential hit*** as 2, and a ***possibly active*** as 1, compounds not identified as definitely toxic are marked as ***probably toxic*** if they score 5 or 6 across the three channels.
4. After removal of toxic compounds, hit compounds can then be ranked either directly by using their end of test fluorescence ratio, or by adding a further step by taking the ratio of this value against the HsDHFR ratio. Both approaches have been used for preliminary studies.
5. All other compounds are considered inactive.

6. Note: the compounds marked up in the two categories of toxicity are probably worth examining further to see if the activity of the test strains is sufficiently different to the Hs strain to warrant a deeper investigation; earlier confirmation studies on PvDHFR hits suggested that some toxic compounds might be of interest if examined at lower concentration.

3.4.3 Experimental and Results

The new method was used by Cambridge to predict compounds that were active against PvDHFR or TbDHFR; the output contained several that had not previously been run through Eve's confirmation screen.

	Cambridge predictions	**Not previously tested by Eve**
PvDHFR	67	12 (TS3 & TS6)
TbDHFR	46	36 (TS4)

Table 3.24: New candidate compounds using Cambridge's filters

Additional confirmation screens were run to evaluate these compounds. The five control compounds were also run in each screen; in the TS6 confirmation screen it was noted that the concentration curve for Pyrimethamine (the main PvDHFR positive control) was weaker than in previous studies, but was still clearly active.

New activity predictions for TbDHFR

All but two of the 46 compounds have now been tested in a confirmation screen.

Predictions for each of the 44 compounds have also been made using Eve's original process; this splits compounds into various groups, including: toxic, sapphire-active (TbDHFR), probably toxic/sapphire hit, inactive, and autofluorescent.

The subsequent confirmation curves are split into *inactive*, *hit*, and *possibly toxic*.

			Confirmation screen results		
			Inactive	**Hit**	**Possibly toxic**
Cambridge active predictions		44	20	7	17
TS4 mass screen prediction	Toxic	18	6	1	11
	Active	11	2	4	5
	Active/probably toxic	4	2	1	1
	Inactive	4	3	1	-
	Autofluorescent	7	7	-	-

Table 3.25: TbDHFR predictions versus confirmation results

New activity predictions for PvDHFR

Of the 67 compounds new predictions, all but one compound has now been tested in a confirmation screen. Again, the original predictions for the mass screen were also compared against these confirmation results

The subsequent confirmation curves are split into *inactive*, *hit*, and *possibly toxic* groups. The final confirmation screen groupings are based on a compilation of the TS3 and TS6 curves.

			Confirmation screen results		
			Inactive	**Hit**	**Possibly toxic**
Cambridge active predictions		66	8	30	28
TS3 mass screen prediction	Toxic	24	3	3	18
	Active	30	-	25	5
	Active/probably toxic	6	2	1	3
	Inactive	5	2	1	2
	Autofluorescent	1	1	-	-
TS6 mass screen prediction	Toxic	15	3	-	12
	Active	32	1	23	8
	Active/probably toxic	12	1	4	7
	Inactive	6	2	3	1
	Autofluorescent	1	1	-	-

Table 3.26: PvDHFR predictions versus confirmation results

Eve's predictions for TS3 & TS6

When applied to the PvDHFR results in TS3 and TS6, Eve predicts 316 (21 probably toxic) and 303 (19 probably toxic) compounds to be active, respectively.

Confirmation screens were initially run on the top candidates from each of these lists.

Eve's top predictions for TS3 PvDHFR

Two confirmation screens were run on TS3 Maybridge Hitfinder compounds; the predictions for these compounds were based on both parasite strains in the screen (Pv & PfRdhfr), and were compiled retrospectively for the compounds in the December 2010 screen as the rules had not been developed at this point. Of the 85 compounds tested in the confirmation screen, 67 were listed as possible hits from the mass screen and 52 were subsequently confirmed.

			Confirmation screen		
			Inactive	**Hit**	**Possibly toxic**
Dec_2010	Toxic	8	-	1	7
	Active	19	1	16	2
	Active/probably toxic	2	-	1	1
	Inactive vs PvDHFR	3	3	-	-
	Autofluorescent	-	-	-	-
CS_ 63_2	Toxic	-	-	-	-
	Active	48	2	36	10
	Active/probably toxic	-	-	-	-
	Inactive vs PvDHFR	4	4	-	-
	Autofluorescent	1	-	-	1

Table 3.27: TS3 Mass screen PvDHFR predictions vs confirmation results

34 of the 66 Cambridge predictions were included in the original TS3 confirmation screens; 27 were *hits*, 7 were *possibly toxic*.

Figure 3.11: (a) Active TS3 compounds (b) Confirmed TS3 hits

Eve's top predictions for TS6 PvDHFR

Similarly, two main confirmation screens (cherrypick & CS_77_7) were run on compounds selected visually after running the TS6 mass screen (Pv & PfDHFR). The results from these screens have been combined to give 75 selections, and Eve's activity predictions have again been applied retrospectively:

			Confirmation screen		
			Inactive	**Hit**	**Possibly toxic**
TS6 mass screen ML	Toxic	2	1	-	1
	Active	48	2	35	11
	Active/probably toxic	14	-	4	10
	Inactive vs PvDHFR	11	5	5	1
	Autofluorescent	-	-	-	-

Table 3.28: TS6 Mass screen PvDHFR predictions vs confirmation results

48 of the original 75 mass selections were predicted to be active using Eve's rules; there was an overlap of 22 compounds between these and the Cambridge predictions.

Figure 3.12: (a) Active TS6 compounds (b) Confirmed TS6 hits

Eve's top predictions for TS4 TbDHFR

53 compounds were selected using TS4 mass screen data as potentially active versus TbDHFR.

		Confirmation screen		
		Inactive	**Hit**	**Possibly toxic**
Predicted as active	53	14	13	26

Table 3.29: TS4 Mass screen TbDHFR predictions vs confirmation results

The new filters predicted 44 active compounds, of which 7 were confirmed; these included three compounds not identified by Eve's rules:

9504: Eve predicts this to be inactive.

3951: Eve predicts this as probably toxic/Tb-active.

12803: Eve predicts this to be toxic.

Eve's rules found 9 hits that were not identified by the new filters.

Figure 3.13: (a) Active TS4 compounds (b) Confirmed TS4 hits

In general, the hits identified for the TbDHFR strain were of much lower strength than those for PvDHFR; this has probably contributed to the lower proportion of confirmed hits identified by either method.

3.4.4 Precision-Recall graphs for PvDHFR predictions

A comparison of how well the two approaches classify compounds can be made using Precision-Recall curves (**Davis and Goadrich, 2006)**. These are generated using quantification of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN).

		Prediction	
		+	-
Actual	+	TP	FN
	-	FP	TN

$$Recall = \frac{TP}{(TP+FN)}$$

$$Precision = \frac{TP}{(TP+FP)}$$

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Figure 3.14: Calculations for precision and recall parameters

Across all three sets of predicted hits for PvDHFR (Cambridge's, Eve's for TS3 and TS6), some 128 different compounds have been identified; each of these compounds has been run in a confirmation screen. These compounds were ranked based on the methods below, and these rankings were used to build P-R graphs:

- When Cambridge's predicted hits were supplied they also included a ranking based on HsDHFR:PvDHFR end of test growth ratios. The Cambridge method only originally predicted 66 active compounds; the remaining 52 were subsequently ranked based on the TS3/TS6 growth data.
- Eve's predicted hits were ranked initially by whether they were identified as a discrete sapphire hit, and secondly by their end of test PvDHFR growth relative to the negative control.

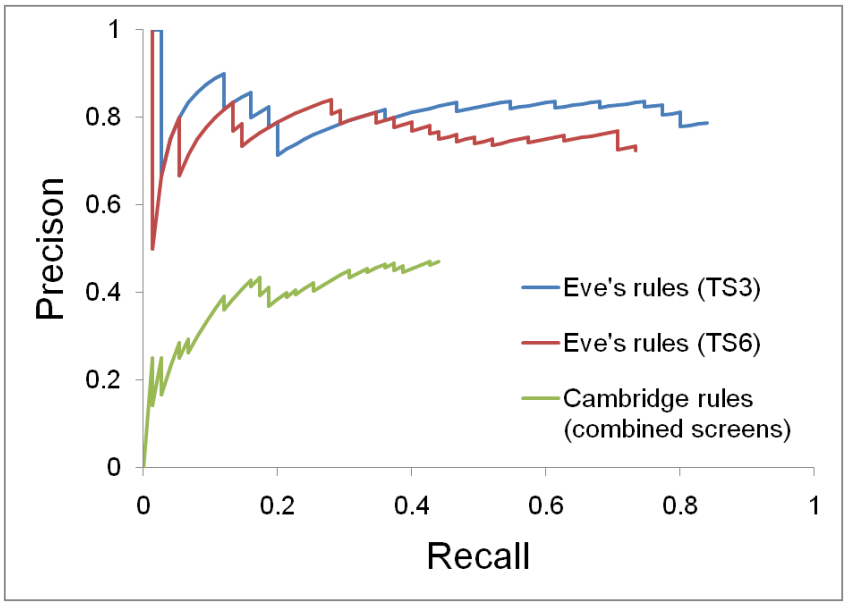


Figure 3.15: Precision-Recall curves for PvDHFR predictions

In addition to a comparison of Cambridge's method with Eve's, the mass screen predictions for TS3 and TS6 could also be cross-compared with the confirmed hits.

TS3 screen data TS6 confirmed hits		Prediction		Recall = 25/(25+10) = 0.71 Precision = 25/(25+5) = 0.83 F = 0.77
		+	-	
Actual	+	25	10	
	-	5	8	

Table 3.30: P-R result for TS3 mass screen data

TS6 screen data TS3 confirmed hits		Prediction		Recall = 31/(31+21) = 0.60 Precision = 31/(31+9) = 0.78 F= 0.67
		+	-	
Actual	+	31	21	
	-	9	6	

Table 3.31: P-R result for TS6 mass screen data

3.4.5 Interpretation of results

Confirmed results

A simple breakdown of the relative success rates for Cambridge's and Eve's filters can be made for PvDHFR:

- 30 of the 66 hits (45%) predicted by Cambridge's filters were confirmed.
- 52 of 67 (78%) were confirmed for Eve's filters for the TS3 mass screen.
- 35 of 48 (73%) were confirmed for Eve's filters for the TS6 mass screen.

Note: evaluation of Eve's mass screen data gives much longer lists of potentially active compounds for PvDHFR than those run in the confirmation screens.

The new prediction filters gave a lower proportion of novel confirmed hits; this effect was seen very strongly in TS3 (3 unique hits from 30 confirmed, versus 25/52 for Eve), and less so in TS6 (16/30 versus 21/35).

Benefits and limitations of Cambridge's filters

The new method does not necessarily remove compounds active versus the Hs strain. Empirical evidence suggests that potential hit compounds that are even only moderately active against Hs at 10 μM have restricted discrimination through the concentration range used in the confirmation screen. This is also evident in the P-R curves, where many of the compounds predicted to be most active were found to be toxic in confirmation screens.

By only using end-of-test growth, the new method is also limited in its ability to identify autofluorescent compounds, and misses variations in growth and toxicity that the lagtime and doubling time can suggest.

The idea of using multiple screens to identify generally toxic and problematic compounds is good, and has yet to be tried on a full scale in Eve's code; earlier work identified many such compounds, but it has yet to be shown whether it is suitable to remove them from evaluations as they may still offer useful information for future Active Learning processes.

The curves below show the possibilities for false negatives when filtering candidate compounds using only the end-of-test growth. Depending upon which target curve is labelled as a false negative, the compound could be categorised as either inactive or as possibly toxic:

False negative Hs → Possibly toxic candidate

False negative parasite → Unidentified active candidate

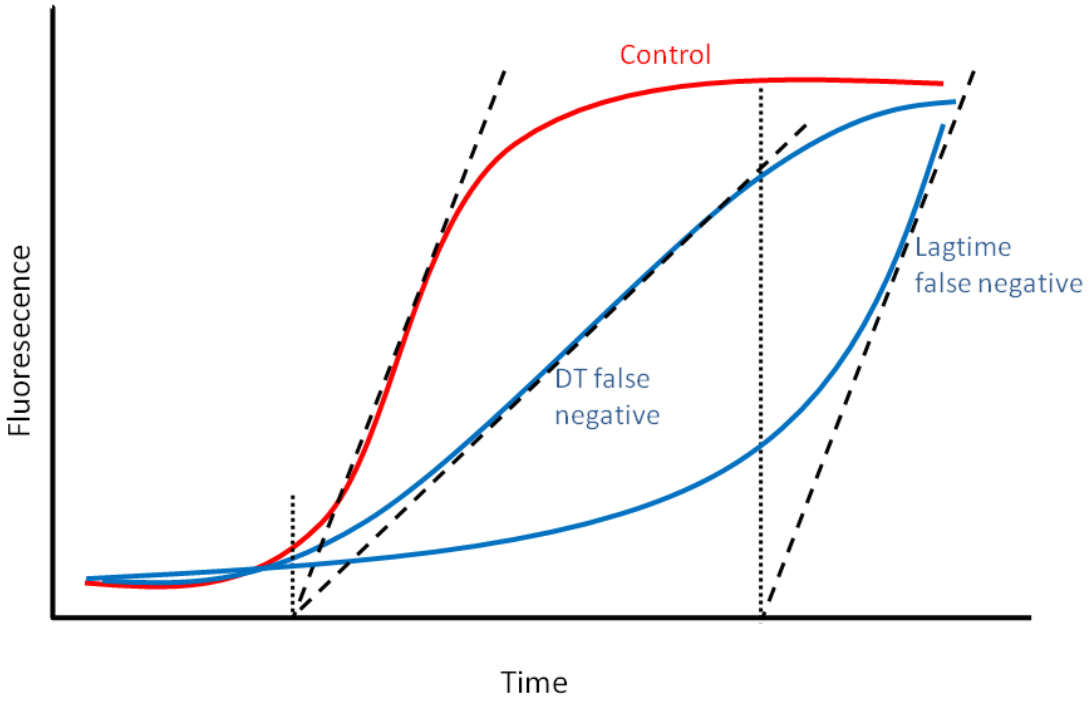


Figure 3.16: Effect of false negatives on candidate classification

3.4.6 Conclusions

The two sets of filters used for selecting hits for Robot Eve use slightly different approaches. The method provided by Cambridge is based solely on end-of-test growth of the parasite yeast strain and its relative performance versus the Hs control; the approach initially developed for Eve uses similar factors to those of this method, but also uses intermediate growth data in the form of lagtime and growth rate.

In the work described here, Eve's decision tree rules clearly generate a higher proportion of confirmed hits than the alternative approach. Whilst the new rule set is simpler and has been developed using cross-screen datasets, it appears to be limited in its ability to eliminate toxic and autofluorescent compounds, and does not select some of the strongly active compounds identified by Eve's code.

3.5 *In vivo* experiments with parasite targets

3.5.1 Validation of confirmed hit compounds by demonstrating their action against *Trypanosoma brucei* in culture

36 hits against yeast strains encoding *T.brucei*, *T.cruzi* or *L.major* targets were selected for validation by the Cambridge School of Biosciences using intact *T.brucei* parasites (**Bilsland *et al.*, 2013**). 18 of the tested compounds were able to kill *T.brucei* Lister 427 bloodstream form parasites at 10 μM (after 48 hours) and 5 additional compounds were responsible for a severely reduced parasite yield (Table 3.32). The drugs capable of killing the parasite at 10 μM were tested in titration experiments to determine the minimum concentration necessary to kill *T.brucei* Lister 427 parasites. All of the 10 μM hits were confirmed and 7 of the compounds showed some effect at 1 μM, 4 were effective at 100 nM and 2 were effective at 10 nM.

When using two *T.brucei* strains for a quantitative study (Lister 427 - a monomorphic laboratory isolate, and EATRO 1125 - a pleomorphic isolate with limited passage history [i.e. being exposed to a relatively low number of subcultures of the strain in order to minimise variations]), it was observed that compounds ID_4584 and ID_14129 could kill virtually all parasites after 48 hours at concentrations as low as 10 nM.

T. brucei target	ID	Lister 427 (growth score)					Lister 427 (% growth after 24/48 h), @ nM			Eatro 1125 (% growth after 24/48 h), @ nM		
		µM		nM			1000	100	10	1000	100	10
		10	1	100	10	1						
DHFR, NMT	3259	L	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
(^{Tc}DHFR)	3951	D	D	S	L	L	0/0	34/7	80/88	22/2	50/18	73/80
DHFR	3978	D	S	L	L	L	40/0.2	79/57	101/64	nt	nt	nt
DHFR	4584	D	D	S	S	L	0/0	5/0.2	36/46	0/0	0/0	14/0.7
DHFR	5422	D	L	L	L	L	nt	nt	nt	nt	nt	nt
DHFR	5833	D	L	L	L	L	nt	nt	nt	nt	nt	nt
DHFR, PGK, NMT	6210	L	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
DHFR	6480	D	L	L	L	L	nt	nt	nt	nt	nt	nt
DHFR	6673	L	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
DHFR	6777	D	L	L	L	L	nt	nt	nt	nt	nt	nt
PGK	7107	L	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
PGK, DHFR	8353	L	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
(^{Lm}DHFR)	9034	L	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
DHFR, PGK, NMT	9504	L	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
NMT	9877	S	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
DHFR	11133	D	L	L	L	L	nt	nt	nt	nt	nt	nt
DHFR	11250	L	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
(^{Tc}DHFR)	11783	D	L	L	L	L	nt	nt	nt	nt	nt	nt
DHFR, PGK, NMT	12135	D	L	L	L	L	nt	nt	nt	nt	nt	nt
(^{Tc}DHFR)	12803	D	D	L	L	L	20/26	89/88	111/88	nt	nt	nt

Table 3.32 (part 1): Hit validation in *Trypanosoma brucei*

T. brucei target	ID	Lister 427 (growth score)					Lister 427 (% growth after 24/48 h), nM			Eatro 1125 (% growth after 24/48 h), nM		
		µM		nM			1000	100	10	1000	100	10
		10	1	100	10	1						
(^{Tc}DHFR, ^{Tc}PGK)	12830	D	L	L	L	L	nt	nt	nt	nt	nt	nt
DHFR	12913	S	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
DHFR, PGK	13015	D	L	L	L	L	nt	nt	nt	nt	nt	nt
DHFR, PGK, NMT	13085	S	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
PGK	13309	S	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
PGK	13528	L	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
DHFR	13692	S	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
DHFR	14129	D	D	D	S	L	0/0	0/0	20/68	0/0	0/0	0/0
NMT	14238	D	D	L	L	L	nt	nt	nt	nt	nt	nt
DHFR	14244	L	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
DHFR	14608	L	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
DHFR	14952	D	L	L	L	L	nt	nt	nt	nt	nt	nt
DHFR	16236	D	D	S	L	L	0/0	32/0	97/83	43/65	65/88	76/113
(^{Tc}DHFR)	16718	L	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
DHFR	16724	L	nt	nt	nt	nt	nt	nt	nt	nt	nt	nt
NMT	17196	D	L	L	L	L	nt	nt	nt	nt	nt	nt

D = dead; S = sick or slow growing; L = live; nt = not tested. Highlighted cells: orange = strong result, yellow = intermediate result.

$$\%\ Growth\ after\ 24/48\ hours = 100 \times \frac{cell\ count\ with\ drug}{cell\ count\ no\ drug\ control}$$

Table 3.32 (part 2): Hit validation in *Trypanosoma brucei*

3.5.2 Extracted *Plasmodium sp*. DHFR results (Mahidol, Thailand)

Table 3.33 summarizes the inhibition by 3 selected JHCCL compounds of purified triple mutant PfDHFR (N51I + C59R + S108N) compared to DHFRs from *P. vivax* and human DHFR.

Satraplatin was found not to inhibit the DHFRs from all 3 sources at concentration 100 µM. However, the delivery of this compound was delayed by several weeks, possibly being held by customs in Bangkok; transportation and storage of this compound should be done under refrigerated conditions, and it is possible that its activity will have degraded prior to conducting the work.

Triclosan, was found to be able to inhibit PfDHFR, PvDHFR, and humanDHFR at similar concentration (with IC_{50} of 50 µM).

TNP-470 was found to be able to inhibit only PvDHFR at IC_{50} 0.16 µM, but not PfDHFR nor human DHFR. Further investigations are required to test TNP-470 against resistant PvDHFR. These studies could highlight the potential importance of TNP-470 as selective inhibitor against PvDHFR.

Compound	IC_{50}		
	P. falciparum **DHFR (triple mutation)**	***P. vivax*** **DHFR (wild-type)**	**Human DHFR**
Satraplatin	No inhibition (≥ 100 µM)	No inhibition (≥ 100 µM)	No inhibition (≥ 100 µM)
Triclosan	No inhibition (≥ 100 µM)	25 µM	1.0 mM
TNP-470	No inhibition (≥ 165 µM)	0.16 µM	No inhibition (≥ 165 µM)

Table 3.33: JHCCL hit validation in Plasmodium sp. by Mahidol University

Chapter 4

Development of active learning algorithms for drug discovery

Adventures with yeast, part 4 of 7: Homegrown garlic & rosemary focaccia

600 grams	*white bread flour*
10 grams	*sea salt*
10 grams	*dried yeast*
80 ml	*olive oil (plus a little extra)*
250 ml	*warm water*
Several sprigs	*rosemary*
8 cloves	*garlic*

Make a dough from the first five ingredients; press it out in e.g. a lasagne tray (8"×10") and leave to rise for one hour in a warm place. Stud the dough deeply with garlic and rosemary. Drizzle with olive oil and broadcast a few flakes of sea salt. Bake in a preheated oven at 200°C for 30 minutes.

Several variations of Active Learning strategies were developed; these were all based on SMILES code fingerprints to describe the chemical compound structure. The seed activity inputs for the prototype *active k-optimisation* method were the numerical growth differences, with activity classifications being used for all other strategies. The limited amount of cherry-pick/confirmation data meant that implementation of these strategies in simulations required that a suitable proxy be found for the multiple concentration measurements that would be made during confirmation HTE; log-weighted replicates of mass screen results were found to be suitable for this work.

Several options for unsupervised clustering were examined; these ideas led to a study to find the likelihood of locating an active compound given the active seed input data. Existing mass screen data was used to explore the thresholds at which the Tanimoto Similarity could give an indication of activity likelihood above the background signals, prior to running multiple simulations as described in Chapter 5.

The strengths and weaknesses of clustering and transfer learning were also explored, and strategies were also developed based on their combination.

Definitions were built to describe rare category compounds based on simple learning mechanisms. Measurement of deficiency was adopted to describe the performance of each AL strategy.

4.1 Cherry-picking using *active k-optimisation*

4.1.1 Method

The prototype Active Learning (AL) method for Robot Eve was provided by Kurt De Grave, based on his *active k-optimisation* strategy (**De Grave *et al*., 2008a)**. The inputs for this method are OpenBabel FP2 fingerprints (using the 0/7 configuration) (**Babel**; **James *et al*., 1995**) for compound SMILES codes (**Weininger, 1988; Weininger *et al*., 1989**) (training set and unknowns), and simple growth differences between target and human strains, i.e.

$$(strain\_yield\_ratio_{Hs} - strain\_yield\_ratio_{target})$$

The method is designed to identify the next best set of compounds to test from a given library, whilst avoiding selection of a compound very similar to any in the training set. The method selects compounds using four different search criteria, in order to provide a diverse approach to the exploration of the chemical space, with selection of the next candidates based on their ability to improve the overall model:

- Compounds close to the decision boundary (the *maximium predicted* strategy)
- Compounds that are expected to be maximal at the lower confidence boundary (the *optimistic* strategy)
- Compounds estimated to provide the *most probable improvement* (MPI), as developed in (**De Grave *et al*., 2008a**)
- Compounds selected using the efficient global optimisation algorithm, EGO (the *maximum predicted improvement* strategy) (**Jones et al., 1998**).

By cycling through these strategies, exploration of the full chemical space of the library is thereby balanced with exploitation of areas most likely to contain active examples.

4.1.2 Implementing experimental Active Learning loops

The data analysis techniques and subsequent rules detailed in Chapter 3 were combined with the following, to enable an autonomous process:

- a method to identify 'hit' compounds (from data acquired either during the AL cherry-pick screens, or during a confirmation screen of potentially active compounds highlighted during a mass screen)
- a method to combine the multiple confirmation/intelligent screen cherry-pick results with the simpler mass screen data, to build the information source for the AL feedback loop.

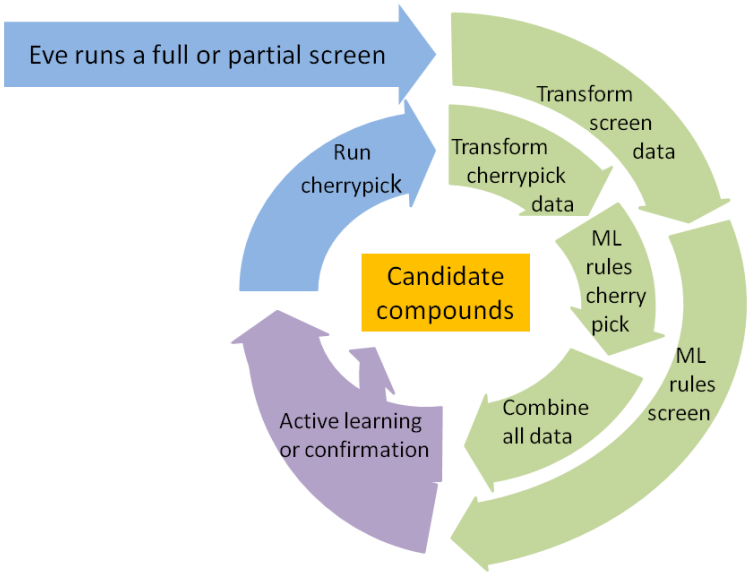


Figure 4.1: Hit compound identification process for Eve

To implement the AL process for Eve, the method is primed by mass screening a section of the library as the training set; the robot runs in this mode until at least x parasite-active compounds have been identified, and the training set will now contain all the compounds on these analysed plates, whether active or otherwise. The AL algorithm is then run iteratively to select n individual compounds to build a batch for the next round of testing. Each group of selections is then tested by Eve in confirmation mode, and the compounds are classified by activity. The SMILES codes and growth data are then fed back into the training set, and a fresh selection is made from the remaining unknown compounds. In practice, $x = 25$ was chosen

as the switching point in simulation work; other values of $x\,\{50, 75, 100\}$ were also examined using the TS6 PvDHFR/HsDHFR data sets (14099 compounds, 303 PvDHFR total hits).

At the time of the physical experimental AL runs (March/April 2011), Eve was capable of running 14 plates in a single batch, and each plate could hold 8 compounds (8 replicates at 5 concentrations, plus positive and negative controls). A choice was made to limit the batch size n to 96 compounds spread over 12 plates.

Eve ran two rounds of selections based on this approach. The first round of 96 compounds (CS_77_3_6_20110325115514.csv) was selected after a partial mass screen (4800 compounds, 69 hits).

For the second round, it was also originally planned to test 96 compounds, but this became extended to 188 compounds (102 in CS_77_4_7_20110404123211.csv, 86 compounds in CS_77_5_9_20110411115111.csv) to examine different methods of weighting the data from the first cherry-pick loop. If the full data set from the cherry-picking loop were used, it was envisaged that this could create two problems: significant bias towards later selections, and a much larger data matrix with accompanying computational difficulties. To examine different approaches to the problem of combining cherry-pick and mass screen data, seven versions of weighted cherry-pick data ($strain\_yield\_ratio_{Hs} - strain\_yield\_ratio_{target}$) were tested in the second loop:

i. all replicates for each compound
ii. all replicates multiplied by $10/conc$
iii. all replicates multiplied by $\log(\frac{10}{conc})$
iv. mean of replicates at each concentration for each compound
v. mean of replicates multiplied by $10/conc$
vi. mean of replicates multiplied by $\log(\frac{10}{conc})$
vii. no additional data (i.e. the next best 96 compounds under loop1 conditions)

The curves for the cherry-pick data were categorised visually, and evaluated against the seven weighting versions (Table 4.1).

	Loop		Weighting options, loop 2						
	1	2	i	ii	iii	iv	v	vi	vii
Compounds	96	188	96	96	96	96	96	96	96
Clean hit	3	9	7	6	8	6	7	9	1
Hit/toxic	-	3	3	3	3	3	3	3	2
Weak activity	3	16	8	8	10	8	8	7	6
Weak/toxic	1	-	-	-	-	-	-	-	-
Possibly toxic	-	8	2	6	3	2	6	2	3
No activity	87	160	76	73	72	77	72	75	84

Table 4.1: Compound activity selected by different cherry-pick weightings

Each weighting method, (i) to (vi), worked well, and far better than simply taking compounds 97-192 from the conditions under loop 1, method (vii). It is suggested that option (vi) would provide an optimised method for combining the cherry-pick data with the mass screen data: the mean result for each concentration is used {5 data points c.f. 40 for options (i) to (iii)} and the behaviour across the concentration range is moderated by the application of a logarithmic weighting.

4.1.3 Active Learning simulations

Ideally this and other algorithms would have been tested physically across the full Maybridge Hitfinder library. However, budget and time constraints required that testing could only be conducted in simulation form.

To follow a similar route to the AL experiments above, the available inputs were:

(i) End-of-test growth data for a group of seed compounds, unclassified.
(ii) A group of “unknown” compounds.
(iii) The SMILES codes for each group.
(iv) Mass screen results and decision tree classifications for the “unknown” group, as a proxy for confirmation results (it was shown in Section 3.4 that the mass screen classification was a good indicator of ground truth for confirmation screen performance).

The seed/loop1/loop2 results from weighting option (vi) were used to predict a third set of 96 compounds, which in turn were evaluated for activity by their mass screen performance. Compound activity for all three loops was compared with other compound loop sets generated with different applications of proxy data. Through

systematic trial it was discovered empirically that a five-fold application of $(strain\_yield\_ratio_{Hs} - strain\_yield\_ratio_{target})$ from the mass screen data set could be used to suitably represent the mean log weighted cherry-pick data from the AL loops. This allowed simulations to be run using Eve's mass screen data, which in turn would allow examination of the overall performance of the *active k-optimisation* method.

Loop	Hits			Weak activity		
	Experiment	Both	Simulation	Experiment	Both	Simulation
1	5	5	5	3	3	5
2	11	10	11	4	4	4
3	4	2	8	0	0	3

Table 4.2: Active compounds in each 96 compound AL loop

Comparison of these *in silico* loops with the original experimental loops showed strong similarity in compound selection. This weighting process has subsequently been used for all simulation work to date.

4.2 Greedy searching as a base case

Whilst the *active k-optimisation* algorithm was designed to use raw data from the partial mass screen, it was also considered possible that selection strategies could be built based on classification data from the same seed compounds. The quantity of input data would be much lower for such strategies due to relatively low numbers of active compounds; this should also allow less complex (and faster) predictions to be computed.

The SimplyGreedy algorithm (Figure 4.2)

SimplyGreedy is designed to simulate finding and testing of the next best n compounds, based on their similarity to existing hits. Candidates are selected from the remaining unknowns, based solely on their proximity to compounds already flagged as active. There are several limitations using this approach, as no consideration is given to either the magnitude of activity, or to the opportunity cost of finding additional compounds around clusters of known activity.

SimplyGreedy has been constructed to give basic predictions of the next best compounds to run, for comparison with a linear, sequential evaluation of the compounds in the library. The inputs are:

(i) A group of seed compounds (classified into *hit*, *probably toxic, toxic, autofluorescent, inactive* as per previous decision tree rules).
(ii) A group of "unknown" compounds.
(iii) The SMILES codes for each group.
(iv) Mass screen results and decision tree classifications for the "unknown" group, as a proxy for confirmation results.

From the seed group of compounds, only those defined as *hits* and *probably toxic* are used, and the rest of the structures are ignored. [Note: later, more complex algorithms also incorporated *toxic* as, if Hs/parasite growth ratio is high, these might yet point to similar structures that are less active versus Hs.]

The SMILES codes for each active seed compound are used to calculate their Tanimoto Similarity (TS) versus the full list of unknown compounds. The resultant matrix (i.e. n active seeds versus the unknown compounds) is sorted so that only the

top TS is retained for each unknown; those compounds with the highest TS are selected from this matrix. [Note: unlike the *active k-optimisation* strategy, this selection process doesn't take the magnitude of activity ($growth_{Hs} - growth_{parasite}$) into consideration; it is effectively a *k*NN selection process where *k* = 1 throughout.]

These new selections are classified using the proxy data, and any hit compounds identified in this loop are compared to the remaining unknowns. The resultant TSs are added to the truncated set from the previous loop, and again the compounds with highest TS are selected after the sorting process.

Loops are run until all unknown compounds have been selected.

Process:

(i) get list of seed hit compounds,
(ii) get list of compounds already tested,
(iii) get full list of hit compounds (from the full mass screen data used to supply the proxy simulation data),
(iv) get list of untested compounds,
(v) build remaining unknown SMILES codes by removing (i) & (ii) from (iv),
(vi) use each "hit" from (i) in turn against the remaining unknowns to select/remove the ones with the top 96 Tanimoto coefficients -- build a list of the selections,
(vii) when the list is complete, e.g. 70 seed hits × 96 selections, sort all by TS and pick the top 96 discrete compounds,
(viii) identify if any from (vii) are in (iii); if so then add to input used at (i) and run loop again from (v). If not then take next 96 from (vii) and so on until the next hit(s) are found.

function SIMPLYGREEDY **returns** list of compound sets to test next and the list of identified hit compound sets for each run of the test

inputs: *k*, the number of compounds per simulation loop
hit, a list of seed hit compounds
unknown, a list of untested compounds
fpt, a list of fingerprints (e.g. SMILES FPT2)
for all library compounds
sim, similarity measure (e.g. Tanimoto similarity) between two compounds based on corresponding fingerprints given in *fpt.*

list_next = Ø; *list_found_hit* = Ø
while *unknown* != Ø **do**:
next = set of *k* compounds in *unknown* with the highest *sim* with a compound in *hit*
add *next* to *list_next*
test compounds in *next***,** store the hit compounds as *found_hit*
hit = hit + found_hit
unknown = unknown – next
add *found_hit* to *list_found_hit*
return *list_next, list_found_hit*

Figure 4.2: The *SimplyGreedy* algorithm

4.3 Alternative Active Learning strategies

4.3.1 Clustering algorithms

It is possible to build clusters of compounds using the SMILES codes and Tanimoto Similarity (TS) coefficients in many ways. The following examples are illustrated in Figure 4.3, and their merits are summarised in Table 4.3.

Example 1

Build clusters containing compounds with $TS > x$ by starting with a single compound a, and adding to it all compounds with $TS > x$ to build a cluster c_a. The next cluster c_b could be built using the compound least similar to a, and the process repeated until all compounds are accounted for.

Each cluster of n compounds could be sampled representatively (e.g. by initially examining $\sqrt[3]{n}$); if any active compounds are found then the priority of this cluster stays high, else the priority is reduced. Compounds from the higher priority clusters are selected for the next round of testing, and these results are used to reprioritise the clusters (thereby allowing some of the lower priority clusters to be circulated back into the list for testing).

x would need to be chosen to ensure there is a minimum size for n, else there might be a large number of orphan/small clusters.

Process:

(i) Pick a start compound a and choose next n_a compounds with $TS > x$ to make cluster c_a

(ii) Make next cluster starting with b (least similar to a); repeat for remaining compounds

(iii) Test & classify $\sqrt[3]{n}$ compounds for each cluster

(iv) Identify priority of clusters for further analyses

(v) Goto (iii)

The initial stage (building the clusters) would be an unsupervised learning process, and then AL would be used to test which clusters to examine preferentially.

Example 2

If a value for x is known which gives a minimum likely threshold for finding a similar hit (as an output of the work using the SimplyGreedy algorithm) then clusters could be built that might be richer in activity, using the known hits as the seed. The non-clustered compounds are then arranged as per Example 1, and sampling of each cluster is conducted. The partial mass screen data would be used to bootstrap the AL loops.

Process:

(i) Get hits from the partial mass screen data.
(ii) Make clusters based on $x_{threshold}$ and each seed hit.
(iii) Make additional clusters as per Example 1.
(iv) Populate clusters with activity data from partial mass screen.
(v) Explore unknowns as in Example 1.

Example 3

Force the clusters to contain n compounds. Run the experiment as in Example 1.

If the initial run is of 1200 out of 14400 compounds, then:

$$N.n = 14400 \text{ and } N\sqrt[3]{n} = 1200 \quad \therefore n \approx 42$$

(Douglas Adams would have been so proud/annoyed/non-plussed*) *delete as appropriate

and $N = 343\ clusters$ (i.e. not dissimilar to the 303 PvDHFR hits).

Example 4

Force the clusters to each contain n compounds, and apply the seed data from a partial mass screen. Continue the experiment as in Example 1.

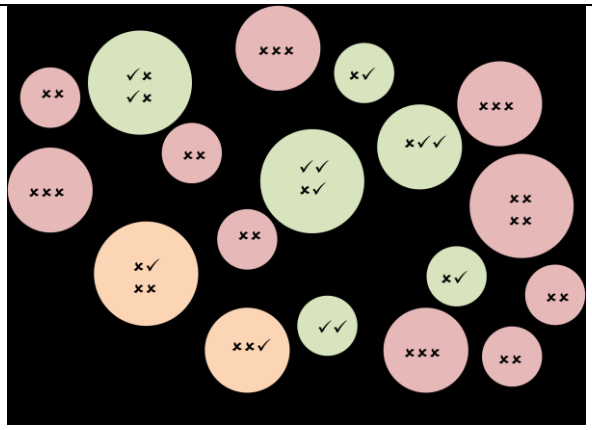
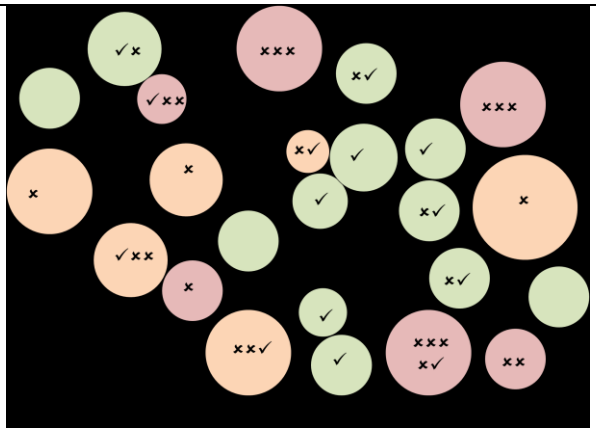
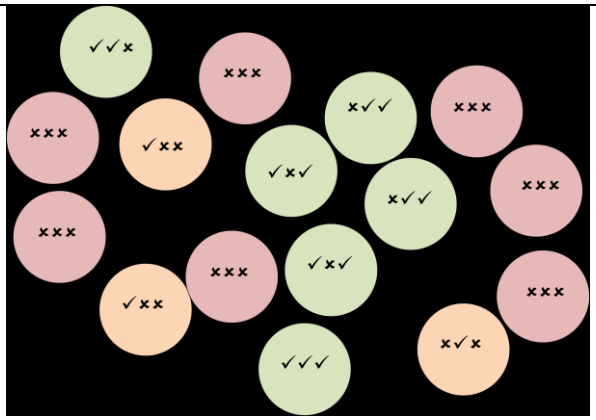
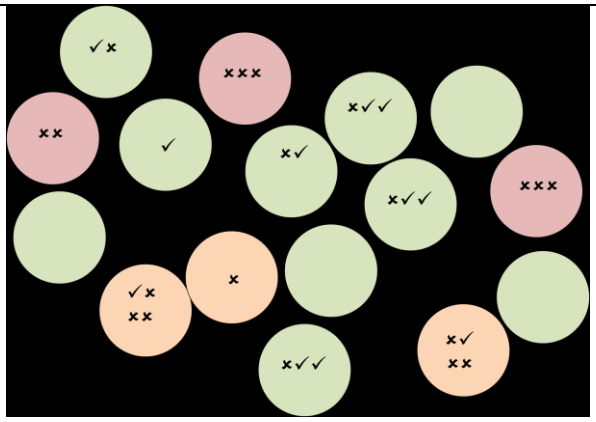
	Example 1: Clusters of various sizes, with $TS > x$ Seeded by representative sampling Continue by priority resampling those with likely activity (Bayesian approach)
	Example 2: Cluster size limited by $TS > x_{threshold}$ Seeded by partial mass screen Priority given to relatively unexplored and active clusters.
	Example 3: Clusters of fixed size Seeded by representative sampling Continue by priority resampling those with likely activity (Bayesian approach)
	Example 4: Clusters of fixed size Seeded by partial mass screen Priority given to relatively unexplored and active clusters.

Figure 4.3: Clusters after the initial evaluation step

Example	Perceived merits	Perceived demerits
1	Early coverage of all clusters.	TS would need to be low ∴ defocused clusters. Possibly too greedy.
2	Clusters based on activity threshold ∴ focused.	Many orphan clusters.
3	Early coverage of all clusters.	Variable TS.
4		Variable TS.

Table 4.3: Merits of example clustering methods

There might also be scope for switching between clustering strategies as a simulation progresses, and as more information becomes available. There are also likely to be benefits to be had by reclustering mid-simulation, e.g. when a proportion of the full unknowns has been evaluated, or when the number of unknowns in a proportion of clusters drops below a threshold. To fully evaluate potential options is beyond the scope of this project, but it should be possible to conduct useful thought experiments using the results from the main simulation algorithms.

4.3.2 Transfer Learning

Cross-screen transfer

If a compound has been identified as active versus one parasite/protein combination, there is a higher prior probability that it might be a hit versus a similar target, e.g. PfDHFR hits applied to PvDHFR targets (**O'Neil *et al.*, 2003).** Alternatively, its presence as a seed might suggest variants in the library that might be better suited to the new target. This prior knowledge can be applied by initially running a batch of previously successful compounds, and then use their freshly tested activity levels to seed the rest of the AL process.

Similarly, compounds that have previously been labelled as autofluorescent or confirmed as cytotoxic could be relegated in the AL process, or even omitted.

This approach is applicable to any of the AL algorithms that are otherwise seeded using a partial mass screen. In practise, the numerical data set used by the *active k-optimisation* would have required complex manipulation for each individual simulation, and would therefore have been difficult to pursue, notwithstanding failed attempts to re-run the original octave-based simulation code on the IBERS HPC. Algorithms based on classified compounds could be more readily tuned to adopt this transfer learning approach.

If simulations indicate that this strategy is advantageous, it would be a fairly simple process to physically build and maintain a seed library based on all previously active compounds. It might also be possible to tailor the seed to the parasite-protein combination based on measurements of previous AL successes.

Internal transfer/multi-objective learning

After some initial proving studies using two strains, the experiments run by Eve all contained three strains. However, the base and prototype AL strategies only use comparisons between the human strain and the target parasite:

SimplyGreedy: parasite-active, ¬Hs-active

active k-optimisation: $(growth_{Hs} - growth_{parasite})$

Therefore, there remains scope to incorporate information from the third strain as there might be compounds active versus this strain that could be lead compounds for the target strain. Similarly, it is conceivable that Hs-active compounds might also be considered as lead compounds for a protein target.

Internal transfer 1: parasite-active, 2nd-parasite-active, ¬Hs-active

Internal transfer 2: parasite-active, 2nd-parasite-active, Hs-active

This technique is not readily accommodated in the *active k-optimisation* strategy, but could be applied to a version based on *SimplyGreedy*.

It was duly suggested that compounds classified as active versus any of the three strains during the course of a simulation could be used to provide multi-objective learning; it would be expected that this would be more advantageous for structurally similar targets (e.g. PfDHFR & PvDHFR), but the effects for less closely related targets were less obvious.

Trans-protein versus trans-species

When attempting cross-screen transfer learning, it is uncertain whether there would be any benefit from using trans-protein (same sp.) seeds e.g. PvNMT & PvPGK to predict PvDHFR. It is assumed that trans-species learning would be more beneficial e.g. PfDHFR & TcDHFR to predict PvDHFR, especially if they were located close together in the phylogenetic tree. The relative performance of each approach could be investigated by multiple simulation iterations using different seed groups.

4.3.3 Rare category detection

One of the more thorny problems for AL is the identification of rare events (**He and Carbonell, 2007**). In the case of finding active drug-like chemicals, it is arguably more valuable to identify rare classes of compounds than those that are similar to known or common scaffolds (**Fischbach and Walsh, 2009; Butler & Buss, 2006**). The use of a greedy, Bayesian algorithm will lead to the more unusual compounds remaining unidentified until late stages of the experiment (Figure 4.4); if an algorithm could be adjusted to allow their earlier detection, it might be feasible to curtail the experiment if sufficient coverage of the library was deemed to have occurred.

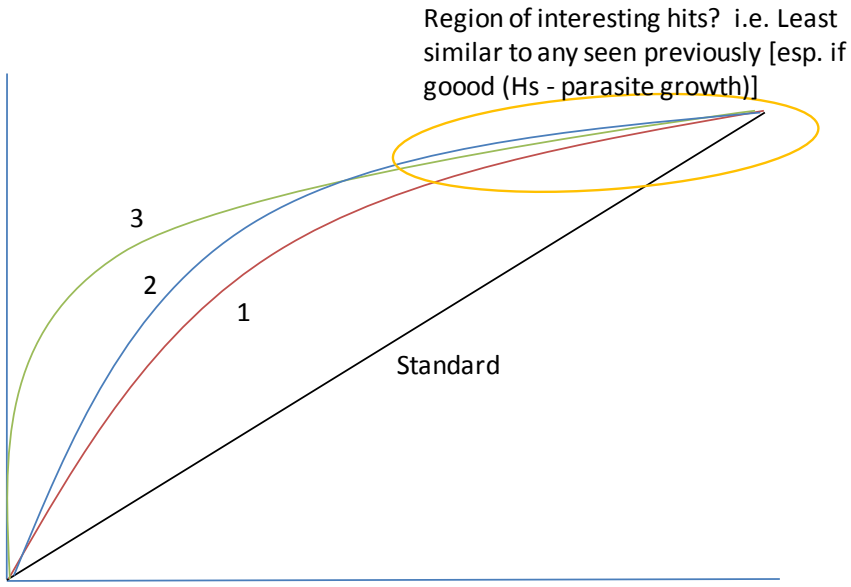


Figure 4.4: Learning curves and rare events

Hierarchical Trees

The clustering algorithms described in 4.3.1 allow variations in: Tanimoto Similarity (TS) for the cluster, number of compounds in the cluster, and various cluster dispersions based on varying TS; these variants can be used to produce different hierarchical trees (**Dasgupta and Hsu, 2008**) which might assist with rare category detection.

Transfer & promotion of prior rare active compounds

Where active compounds from previous experiments have been found to be from small clusters/low populations/rare events, it would be possible to use these compounds preferentially at the start of the next process. It is likely that any hits from well populated active clusters will be found fairly quickly, so this approach might increase the chances of identifying otherwise rare compounds.

Inactive seed compounds as indicators

Any group of seed compounds, whether from an unselective approach such as a partial mass screen, or from a carefully chosen transfer learning set, will invariably contain a large proportion that are inactive against the targets in this new experiment. In the same way that active compounds indicate the potential of similar molecules, so these inactive compounds can be used as priors to indicate potential inactivity for otherwise unknown molecules. Although the changes to prior probabilities are small, similarity to a large number of known inactives might be a good indicator of likely inactivity in the unknown compound.

Imparting this knowledge in a clustering algorithm will allow separation of unknown compounds into three phases:

(i) Those with proximity to active compounds.
(ii) Those with proximity to inactive compounds only.
(iii) Those in their own chemical space.

This approach would give two new ways of ordering the compounds for test. The compounds in section (iii) are likely to contain a higher proportion of rare category compounds, and it would therefore be of benefit to examine these at an early stage. Also, the compounds identified as being similar to inactive compounds could be ranked in terms of the proportion of known inactive compounds in their cluster.

4.4 Activity prediction by chemical structure analysis

4.4.1 Background

There seems little information in the literature concerning a threshold where the TS/activity relationship tails off. When the TS threshold question is discussed in (**Keiser *et al.*, 2007**), it is concluded that there are no hard and fast rules, and the authors choose a threshold of 0.57 based on ligand group data in their work, and also suggest that TS in the range 0.2 - 0.3 shows insubstantial similarity between compounds (note: my work in this section suggests that this relationship tails off a slightly lower limit, closer to $TS \cong 0.45$).

4.4.2 Activity likelihood prediction

There was a need to understand the likelihood that an active seed compound could predict another active compound, within defined similarity limits.

The dataset for TS6 was used for this exercise, with initial questions concerning the likelihoods that:

Hit begats hit.

Toxic begats toxic.

Hs/cherry begats Hs/cherry.

Inactive begats inactive.

The background level for the TS6 PvDHFR_sapphire data set was 306 active compounds in a library of 14099, i.e. 2.17% were deemed active.

Three limits $[TS > (0.40,0.50,0.60)]$were used to build clusters as suggested in section 4.3.1, Example 1. The compounds identified as active in the mass screen were overlaid, and their distribution with respect to cluster size determined.

The size and richness of the clusters were defined as follows:

- Singleton: a single active compound; no other compounds within the TS limit.
- Sparse doubleton: a cluster made up of a pair of compounds within the TS limit, with one active and the other inactive.

- Sparse large: one active compound amongst many inactive compounds within the TS limit.
- Rich: more than one active compound in a cluster containing more than two compounds.

Active seed	Activity prediction	Actives identified	Singleton	Sparse doubleton	Sparse large	Rich
sapphire (306)	sapphire	306	28	44	116	118
	cherry	72	1	10	46	15
	venus	138	8	14	78	38
venus (188)	venus	188	14	29	91	54
	sapphire	88	7	17	35	29
	cherry	79	4	10	46	19
cherry (119)	cherry	119	13	13	57	36
	sapphire	34	1	3	20	10
	venus	80	5	11	40	24
toxic (94)	toxic	94	9	14	32	39
fluoro (56)	fluoro	56	8	-	39	9

Table 4.4: Activity/cluster relationships at $TS > 0.50$

These data show that an active compound is more likely to predict another active against the same target rather than a different target, e.g.

$$p(sapphire\_active, rich\_sapphire\_cluster) \cong \frac{118}{306} = 0.386$$

and that there are quite high proportions of active compounds that occupy a unique (or very sparse) chemical space:

$$p(singleton \vee sparse\ doubleton)_{sapphire} \cong \frac{28 + 44}{306} = 0.235$$

The proportion of compounds that occur as a single active in a large cluster $(n > 2)$ make up the remainder:

$$p(sparse\ large\ cluster)_{sapphire} \cong \frac{116}{306} = 0.379$$

The data also show that an active seed compound from any target is better than a random seed $(p = 0.0217)$, e.g.

$$p(sapphire\_active, rich\_cherry\_cluster) \cong \frac{15}{72} = 0.208$$

$$p(sapphire\_active, rich\_sapphire\_cluster) \cong \frac{118}{306} = 0.386$$

$$p(sapphire\_active, rich\_cherry\_cluster) \cong \frac{15}{72} = 0.208$$

$$p(sapphire\_active, rich\_venus\_cluster) \cong \frac{38}{138} = 0.275$$

$$p(venus\_active, rich\_venus\_cluster) \cong \frac{54}{188} = 0.287$$

$$p(venus\_active, rich\_sapphire\_cluster) \cong \frac{29}{88} = 0.329$$

$$p(venus\_active, rich\_cherry\_cluster) \cong \frac{19}{79} = 0.231$$

$$p(cherry\_active, rich\_cherry\_cluster) \cong \frac{36}{119} = 0.303$$

$$p(cherry\_active, rich\_sapphire\_cluster) \cong \frac{10}{34} = 0.294$$

$$p(cherry\_active, rich\_venus\_cluster) \cong \frac{24}{80} = 0.300$$

$$p(toxic, rich\_toxic\_cluster) \cong \frac{39}{94} = 0.425$$

Figure 4.5: Demonstration of success rates for predicting activity in clusters where $TS > 0.50$, based on data in Table 4.4

Active seed	Activity prediction	Actives identified	Singleton	Sparse doubleton	Sparse large	Rich
sapphire (306)	sapphire	306	4*	7*	93	202
	cherry	136	0	1	90	45
	venus	197	2	3	88	104
venus (188)	venus	188	2	5	73	108
	sapphire	122	2	2	52	66
	cherry	112	0	1	56	55
cherry (119)	cherry	119	3	3	46	67
	sapphire	59	0	0	31	28
	venus	89	0	1	31	57
toxic (94)	toxic	94	0	1	30	63
fluoro (56)	fluoro	56	0	0	36	20

*Active compounds in unique/very sparse chemical space: 3.6% for PvDHFR_sap

Table 4.5: Activity/cluster relationships at $TS > 0.40$

Active seed	Activity prediction	Actives identified	Singleton	Sparse doubleton	Sparse large	Rich
sapphire (306)	sapphire	306	141*	63*	64	38
	cherry	42	14	12	13	3
	venus	106	48	21	24	13
venus (188)	venus	188	84	44	31	29
	sapphire	84	37	23	14	10
	cherry	69	29	21	14	5
cherry (119)	cherry	119	53	26	26	14
	sapphire	25	6	7	9	3
	venus	74	31	18	17	8
toxic (94)	toxic	94	36	29	21	8
fluoro (56)	fluoro	56	20	16	18	2

*Active compounds in unique/very sparse chemical space: 66.7% for PvDHFR_sap

Table 4.6: Activity/cluster relationships at $TS > 0.60$

	Activity predictions		
	$TS > 0.40$	$TS > 0.50$	$TS > 0.60$
$p(sapphire\_active, rich\_sapphire\_cluster)$	0.660	0.386	0.124
$p(sapphire\_active, rich\_cherry\_cluster)$	0.331	0.208	0.071
$p(sapphire\_active, rich\_venus\_cluster)$	0.528	0.275	0.123
$p(singleton \vee sparse\ doubleton)_{sapp\ hire}$	0.036	0.235	0.667
$p(venus\_active, rich\_venus\_cluster)$	0.574	0.287	0.154
$p(venus\_active, rich\_sapphire\_cluster)$	0.541	0.329	0.119
$p(venus\_active, rich\_cherry\_cluster)$	0.491	0.241	0.072
$p(cherry\_active, rich\_cherry\_cluster)$	0.563	0.303	0.118
$p(cherry\_active, rich\_sapphire\_cluster)$	0.475	0.294	0.120
$p(cherry\_active, rich\_venus\_cluster)$	0.640	0.300	0.108
$p(toxic, rich\_toxic\_cluster)$	0.670	0.415	0.085

Table 4.7: Activity predictions at different TS limits

	$TS > 0.40$	$TS > 0.50$	$TS > 0.60$
Clusters(all)	2690	5681	9437
$n = 1$	741	2604	6523
$n = 2$	480	1305	1927
$n > 2$	1469	1772	987
Compounds(all)	14105	14105	14105
$n = 1$	741*	2604*	6523*
$n = 2$	960*	2610*	3854*
$n > 2$	12404	8891	3728

Table 4.8: Cluster sizes at different TS limits

Overall, these data show there is no advantage in isolating singletons & doubletons to search for rare compounds, as the hit rate is not better than the background level:

$$\frac{7+4}{741+960} = \ 0.006\ for\ TS > 0.40\text{:}\quad 0.014\ for\ TS > 0.50;\quad 0.0198\ for\ TS > 0.60$$

Therefore, this tactic would only be suitable for a library where the singletons and doubletons represented a low proportion; at such a point a case could be made to rapidly explore these compounds in pursuit of rare category detection.

4.4.3 Rich and inactive cluster analysis

Rich, active clusters

The active compounds that occur in the rich clusters can be used to determine the threshold at which TS infers likely activity. Using the total number of compounds in rich clusters at each of the three values of TS:

TS	Active compounds	Rich cluster compounds	Section success rate, %	Running success rate, %
> 0.60	38	1058	3.59	3.59
0.50 – 0.60	+80	1494	5.35	4.62
0.40 – 0.50	+84	4076	2.06	3.05
≤ 0.40	+101	other(7477)	1.35	2.17

Table 4.9: Analysis of rich clusters

Having $TS > 0.50$ between compounds gives a good rate of identifying another active compound. Dropping the level to $TS > 0.40$ allows a larger number of active compounds to be found, whilst having a section success rate similar to the background rate and maintain a good running success rate.

Inactive clusters

By extension, the inactive seed compounds that occur in a given cluster might be useful as an indicator for possible inactivity of similar candidate compounds. For larger cluster sizes, an increasing number of prior inactive compounds might reinforce the notion that all similar compounds are inactive, allowing such a grouping to be relegated to a later point in the experiment.

4.5 Active Learning algorithms

4.5.1 General

For each simulation, the relevant full mass screen dataset was mined to provide input data for the simulations. After the initial data analysis step, all results were recorded in the file "full_activity_results_output.csv". All active compounds were identified and stored in separate .csv files (e.g. "full_toxic_compounds.csv", "full_sapphire_hits.csv").

In addition to the prototype *active k-optimisation* and baseline *SimplyGreedy* methods, the ideas in section 4.3 based on TS clustering were resolved into three further processes; these used the evidence for TS threshold-activity relationships recorded in section 4.4 to limit the diversity of seeded clusters. These ideas were aimed at promoting rare category detection, and to determine any beneficial effects of in-screen (endogenous) and cross-screen (exogenous) transfer learning.

It was noted that using TS to create unseeded clusters (as an unsupervised learning approach) would result in potentially more diverse/dilute clusters when operating at similar TS thresholds to the above processes. When active compounds are identified in such clusters, to guarantee that their proximity to other compounds in the cluster is within the threshold-activity boundary, this limit would now be $\sqrt{TS}$.

4.5.2 *Active k-optimisation*

Process:

(i) Select a subset of a full mass screen as a proxy for a partial mass screen. Store this as the seed list of compounds already examined.
(ii) Identify active seed compounds.
(iii) Build a list of remaining unknown compound SMILES codes by removing the seed compounds from the full mass screen list.
(iv) Generate OpenBabel FP2 fingerprints for seed and unknown compounds.
(v) Build an indexed matrix for the activity of the seed compounds, based on: $(strain\_yield\_ratio_{Hs} - strain\_yield\_ratio_{target})$.
(vi) Select next best compounds for examination based on strategy.
(vii) Evaluate; add results to seed compounds and loop back to (ii).

4.5.3 *SimplyGreedy* (see Figure 4.2)

Process:

(i) Select a subset of a full mass screen as a proxy for a partial mass screen. Store this as the seed list of compounds already examined.
(ii) Build a list of active seeds from this subset (e.g. “sapphire_hits.csv”).
(iii) Build a list of remaining unknown compound SMILES codes by removing the seed compounds from the full mass screen list.
(iv) Generate FP2 fingerprints for seed and unknown compounds.
(v) Use each active seed in turn to find the compounds with the highest TS; select the top 96 compounds for each seed.
(vi) When the list is complete, e.g. 70 seeds × 96 selections, sort by TS and select the top 96 discrete compounds.
(vii) Identify if any of these 96 are in the original active compound list. If so, add to seed list and loop back to (iv); if not, return to step (v) and select the next best 96 compounds.

Options for exploring effect of *SimplyGreedy* inputs

The *SimplyGreedy* code allows several variations to be examined:

(a) The number of seed hits can be varied by simulating different partial mass screens.

(b) The active seed input can be changed to include compounds active versus other targets, or “probably toxic” and interesting “toxic” compounds, which in turn will create an internal transfer learning algorithm.

(c) The SMILES/Tanimoto fingerprint type can be changed; this is set at chain length depth search of 7 by default.

(d) The fingerprint type can be changed from “standard” to “extended”; this will then take ring structures into account.

(e) AL curves could be built based on seed hits only; this would help to quantify the benefits of AL.

Note: The changes/variants of the fingerprint are not possible when using OpenBabel, which was a restriction on the coded *active k-optimisation* simulations. Therefore, no direct comparison could be made with this prototype method.

4.5.4 Pre-clustering to induce promotion of rare categories (Figure 4.6)

This algorithm is designed to run in three phases: initially find all compounds that are chemically similar to active seed compounds, then to identify and explore all remaining compounds not tagged as similar to inactive compounds, and finally to evaluate those similar to the inactives. As with all the algorithms (with the exception of *active k-optimisation*) the seed compounds and internal transfer learning can be set to include all streams of activity across the Hs & parasite strains.

In practise, three levels of $TS > x$ were tested: $x = \{0.40, 0.45, 0.50\}$

(i) Select a subset of a full mass screen as a proxy for a partial mass screen. Store this as the list of compounds already examined.
(ii) Build a list of active seeds from this subset.
(iii) Build a list of remaining unknown compound SMILES codes by removing the seed compounds from the full mass screen list.
(iv) Generate FP2 fingerprints for seed and unknown compounds.
(v) Use each active seed in turn to find the compounds with $TS > x$. Build these into a candidate list.
(vi) Select and evaluate the top 96 discrete compounds; categorise them in terms of activity, toxicity or inactivity.
(vii) Add active compounds to seed list and loop back to (iv) until < 96 compounds are identified with $TS > x$. Switch to evaluate Rare Category molecules.
(viii) Cluster the remaining unknown compounds with respect to the inactive compounds at $TS > x$. Split into two sets: $unknown_{inactive}$ and $unknown_{unknown}$. Rank the $unknown_{inactive}$ set in terms of the number of inactive compounds to which they have similarity.
(ix) Evaluate the $unknown_{unknown}$ compounds using the SimplyGreedy algorithm.
(x) Evaluate the $unknown_{inactive}$ compounds in ranked order (fewest similar inactive compounds first).

function PRECLUSTERING **returns** lists of compound set to test next over three phases, and the list of hit compound sets identified in each test loop.

inputs: *k*, the number of compounds per test loop
hit, a list of seed hit compounds
not_hit, a list of compounds tested as non-hit
unknown, a list of untested compounds
fpt, a list of fingerprints for all library compounds
sim, similarity measure between two compounds using *fpt*
threshold, similarity threshold

list_next = Ø; *list_found_hit* = Ø; *inPhase1* =True;
while *unknown* != Ø **do**:

if *inPhase1* **# PHASE 1**

pool = set of compounds from *unknown* whose values of *sim* with a compound in *hit* > *threshold*

if length(*pool*) >= k

next = set of k compounds from *pool* with the highest sim values with a compound in *hit*

test compounds in *next*, store the hit compounds as *found_hit* and non-active compounds as *found_nothit*

update *hit, not_hit, unknown, list_next, list_found_hit* accordingly

else

store all compounds to set *unknown3* if their values of *sim* with a compound in *not_hit* > *threshold*; otherwise store them to set *unknown2*

inPhase1 = False # phase switch

if (not *inPhase1* and *unknown_phase2* != Ø) **# Phase 2**

call function SIMPLYGREEDY(**input**: unknown=*unknown2)*

update *list_next* and *list_found_hit, non_hit accordingly*

if (not *inPhase1* and *unknown3* != Ø) **# Phase 3**

for each compound in *unknown3*, calculate *sim* value with every compound in *non_hit*; store the number of *sim* values > threshold to *nothit_counts*

next = *k* compounds in *unknown3* with the lowest *nothit_counts*

test compounds in *next*, store the hit compounds as *found_hit* and non-active compounds as *found_nothit*

update *hit, not_hit, unknown, list_next, list_found_hit* accordingly

return *list_next, list_found_hit*

Figure 4.6: The *preclustering* algorithm

4.5.5 Transfer Learning from other parasites (Figure 4.7)

Rather than running a partial mass screen to provide active seeds, this algorithm selects a tranche from previously completed screens to use as the seed.

(i) Identify previously evaluated parasite/strain combinations to be used as a transfer learning seed.

(ii) Build a seed subset based on the full list of target-active compounds.

(iii) Evaluate this seed subset as a proxy for a partial mass screen. Store this as the list of compounds already examined.

(iv) Continue from step (ii) of *SimplyGreedy*.

function TRANSFERLEARNING **returns** lists of compound sets to test next, and the list of hit compounds identified for each test.

inputs: *transfer*, a list of seed hit compounds tested with a different strain
k, the number of compounds per loop
hit, a list of seed hit compounds in the transfer screen
unknown, a list of untested compounds, initially the full library
fpt, a list of fingerprints for all library compounds
sim, similarity measured between two compounds using *fpt*

list_next = Ø; *list_found_hit* = Ø

test each compound in *transfer*, store the set of hit compounds to *transfer* in *found_hit*

hit = hit + found_hit

unknown = unknown - transfer

if *unknown != Ø*

call SIMPLYGREEDY (*k, hit, unknown*) and update *list_next, list_found_hit*

return *list_next, list_found_hit*

Figure 4.7: The TransferLearning algorithm

4.5.6 Transfer Learning with *Preclustering* (Figure 4.8)

The ideas for preclustering and transfer learning are brought together in the expectation that the benefits accrued by each method will be cumulative and build a more efficient selection process.

(i) Identify previously evaluated parasite/strain combinations to be used as a transfer learning seed.

(ii) Build a seed subset based on the full list of target-active compounds.

(iii) Evaluate this seed subset as a proxy for a partial mass screen. Store this as the list of compounds already examined.

(iv) Continue from step (ii) of *Preclustering*.

function TRANSFERLEARNING WITH PRECLUSTERING **returns** lists of compounds to test next, and the lists of hit compound sets for each test loop

inputs: *transfer*, a list of seed hit compounds versus a different strain
k, the number of compounds per loop
hit, a list of seed hit compounds in the transfer screen
not_hit, a list of compounds tested as not-hit
unknown, a list of untested compounds, initially the full library
fpt, a list of fingerprints for all library compounds
sim, similarity measure between two compounds

list_*next* = Ø; list_*found_hit* = Ø

test each compound in *transfer*, store the set of hit compounds in *transfer* to *found_hit,* and non-hit compounds to *found_nothit*

hit = hit + found_hit;

not_hit = not_hit + found_nothit;

unknown = unknown - transfer

if *unknown != Ø*

call PRECLUSTERING (*k, hit,not_hit, unknown*) and update *list_next, list_found_hit*

return *list_next, list_found_hit*

Figure 4.8: The *TransferLearning with preclustering* algorithm

4.6 Simulations, methods and evaluation techniques

4.6.1 Datasets for simulations

All simulation datasets used in the learning algorithms were derived from the original mass screen experiments. The decision tree activity results were taken as absolute, and acted as a proxy for confirmation results.

The learning algorithms based on a simulated partial mass screen used the raw data from a series of mass screen plates as the seed group; these were selected to provide the required minimum number of active seeds (typically $n \geq 25$). Several simulations were run for each target, with a different batch of seed compounds selected from the mass screen in each case.

The other learning algorithms used a seed compound list defined by previous screen results as a transfer learning (TL) process. In these instances the number of active seeds was therefore not predetermined, and depended on the response of the new parasite target to compounds with high potential. Only one simulation could be run for each transfer seed/target combination.

The partial mass screen data sets used in the *active k-optimisation* and *SimplyGreedy*/preclustering simulations were slightly different: the RCDK SMILES fingerprint tool (**Guha, 2007**) for R didn't support work with chiral centres, so all such molecules had to be screened out at the start of the simulation process. This did not affect the *active k-optimisation* simulations (where the SMILES strings were parsed by OpenBabel) but meant that 164 compounds could not be evaluated in the *SimplyGreedy* & preclustering simulations; none of these were identified as hits in any mass screen, so the effect on the simulation results is expected to be negligible.

The seed data for *active k-optimisation* was used in the form:

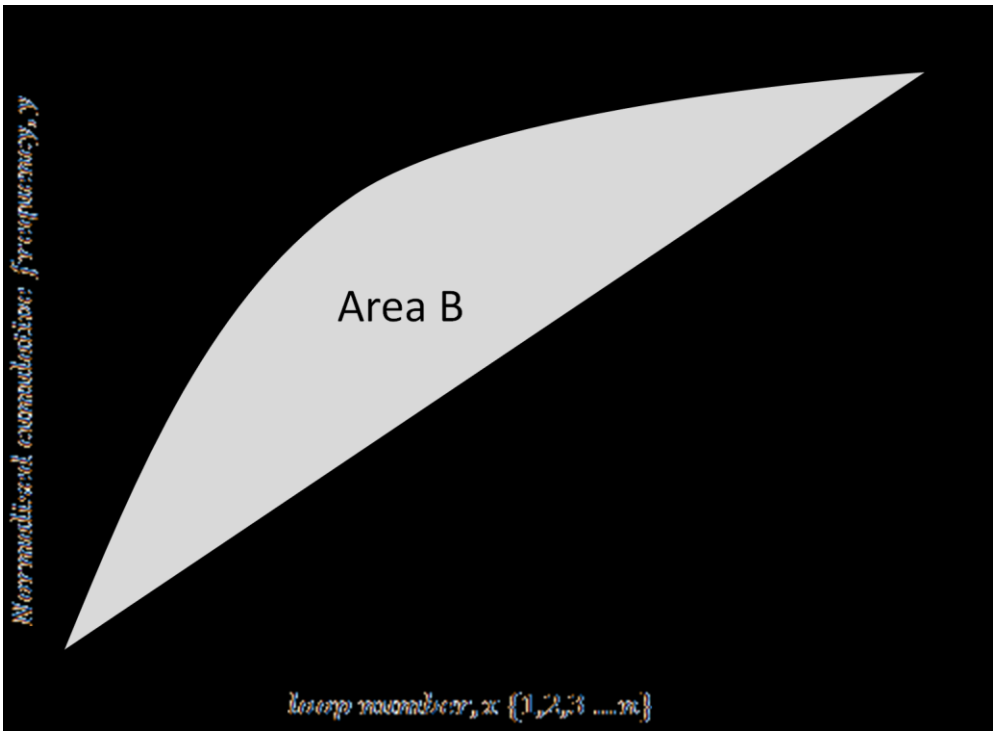
$$(strain\_yield\_ratio_{Hs} - strain\_yield\_ratio_{target})$$

This dataset expanded with each additional evaluated compound.

The seed data for all other simulations was applied as a list of active compounds, and this set was incremented each time a new active compound was discovered.

4.6.2 Deficiency measurement of general/rare category curves

The effectiveness of the AL algorithm can be measured by comparing the yield curve with a baseline curve to give a deficiency measurement (**Baram *et al.*, 2004**). For the AL simulations, the baseline curve is defined by the linear rate at which active compounds would be found if no learning was occurring. It is also feasible to compare any pair of AL curves by similar measurements.



$$Area\ A = \frac{\sum_1^n (1 - y)}{n} \qquad Area\ B = \frac{\sum_1^n (\frac{x}{n})}{n}$$

$$Deficiency = \frac{Area\ A}{Area\ A + Area\ B}$$

Figure 4.9: Comparing learning curves by deficiency measurements

Rare category detection

It is reasonable to assume that a greedy selection mechanism based on TS will make it difficult to find active compounds that have low similarity to other active seeds. If these orphan compounds are in sparsely populated regions of the chemical space then it is highly likely that they will not be detected until the late stages of a simple search e.g. *SimplyGreedy*. If the last 5%/10% of active compounds found by *SimplyGreedy* are taken as a baseline for rare category detection, deficiency measurements can be made by comparing their occurrence in other AL simulations based on the same seed/unknown datasets.

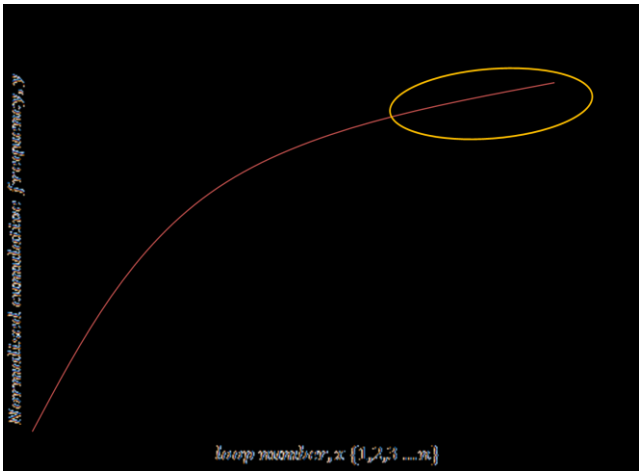


Figure 4.10: Definition of rare category compounds in SimplyGreedy

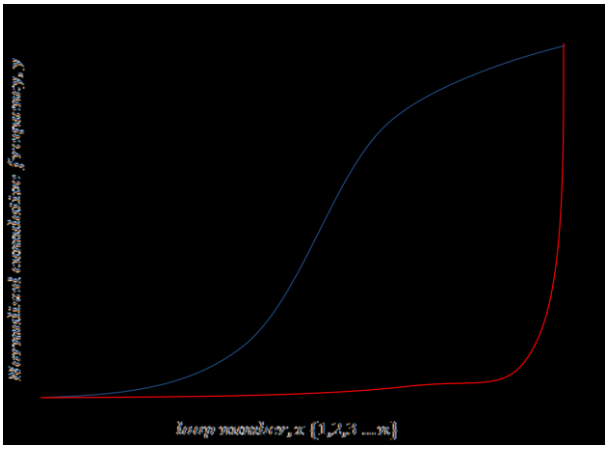


Figure 4.11: Example rare category deficiency curves for preclustering/SimplyGreedy

Chapter 5

Simulations of Active Learning for drug discovery

*Adventures with yeast, part 5 of 7: Aberystwyth sourdough bread**

~1.5 kg	*mixture of white and wholegrain (e.g. rye) bread flours*
~1.5 kg	*water*
10 grams	*sea salt*

Day 1: Mix 100 grams of flour with 100 grams of water in a large pyrex bowl. Take the mixture on an open air walk along Aberystwyth prom, then cover it with clingfilm and leave it in a warm place for 24 hours.

Day2: Mix in 100 grams of flour, and sufficient water to keep the mixture sloppy. Cover tightly and store at room temperature.

Day 3 to day 7: Each day, discard half the mixture, then mix in a further 100 grams of flour and water. Cover tightly and store at room temperature.

The starter can then be maintained, with samples extracted from it for making bread:

Add 150 grams of starter to 250 grams of flour and 250 ml of water. Cover and leave overnight. The following day, add 300 grams of flour and 10 grams of salt, and mix into a stiff dough. Knead the dough and then leave it to rise for a couple of hours (a more sedate process than ordinary bread making). Knock it down, knead again, then allow to prove for another couple of hours in a proving basket.

Bake at 230°C for 40 to 45 minutes.

** In deference to San Francisco sourdough, that other west coast staple.*

This chapter describes how five discrete Active Learning strategies were applied to the mass screen data to simulate high throughput experimentation for drug discovery.

As seen in section 4.1.2, only limited cherry-picking work could be conducted due to resource restraints; the *active k-optimisation* algorithm was evaluated physically, but only for the very early phase of the cherry-picking screen for PvDHFR.

The proxy confirmation screen data derived in section 4.1.3 was subsequently used to portray potential activity, and allowed the alternative Active Learning algorithms to be examined and compared *in silico*. The proxy confirmation data for all targets was used individually in *active k-optimisation*, *SimplyGreedy* and *preclustering* simulations, and data from combinations of targets were used to examine the transfer learning ideas.

The settings for the simulations were chosen to examine a wide range of the features of these algorithms, whilst endeavouring to keep a consistent approach. In general, the seed group of compounds were maintained across those algorithms without a transfer learning element; full details of the set up are embedded in each of the algorithm analysis sections.

Using the deficiency measurement described in section 4.6.2, this chapter examines the absolute and relative ability of both the prototype AL algorithm and those developed in Chapter 4 to select active compounds. The *SimplyGreedy* approach is compared to the linear results expected of random screening, and the other algorithms are then compared to each of these as baselines. The ability of each algorithm to search the chemical space and identify rare category compounds (as defined in their selection by *SimplyGreedy*) is similarly evaluated.

5.1 General examples of learning curves (TcDHFR simulations)

General curve descriptions

Normalised hit cumulative frequency curves (in red, below) display the algorithm's identification rate for active compounds compared to the expected stochastic discovery rate (in black); these are used for deficiency calculations. The variant for the performance of preclustering algorithms (right hand graph) also include two vertical lines which identify where the algorithm switches between strategies.

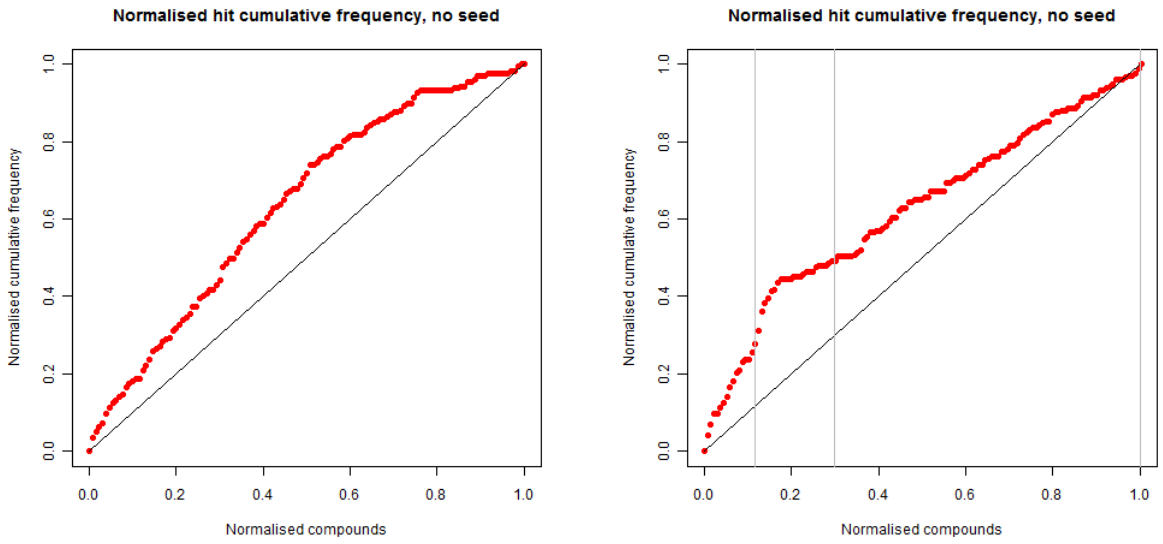


Figure 5.1: Examples of normalised hit cumulative frequency curves

Normalised actives cumulative frequency curves (Figure 5.2) are an extension of the above; compound activity versus the chosen target (red), the second parasite target (green), human strain (blue), and likely general toxicity (orange) are depicted.

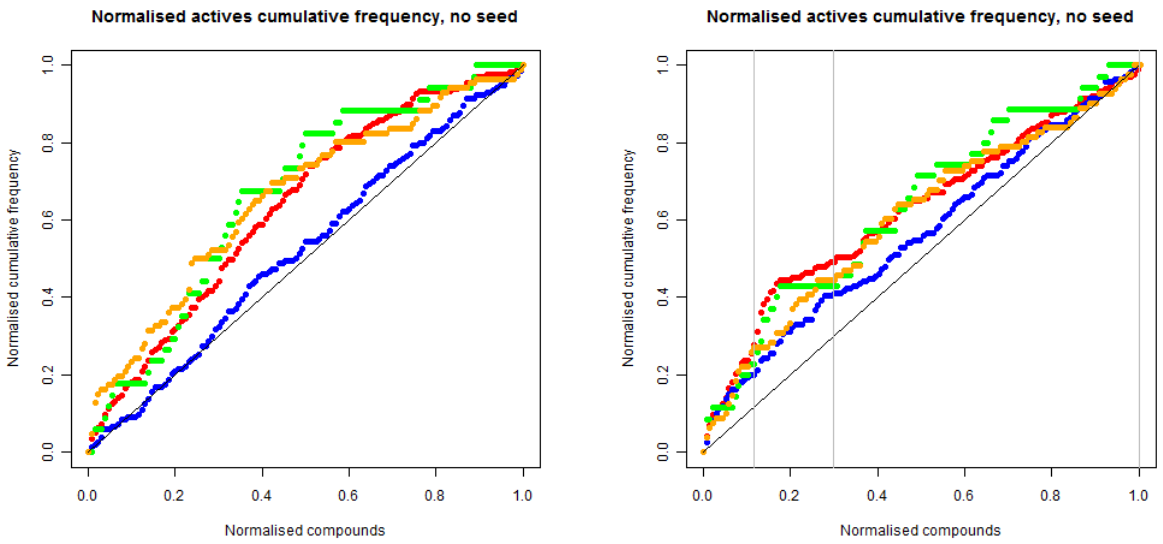


Figure 5.2: Examples of normalised actives cumulative frequency curves

Deficiency curves for rare category compounds

These display the rate at which the algorithm under test (the black curve in this example) identifies rare compounds. In the example below, the rare compounds are defined as the last 5% of active compounds found by the *SimplyGreedy* stategy (red curve). Again, deficiency measurements are calculated by comparing these curves.

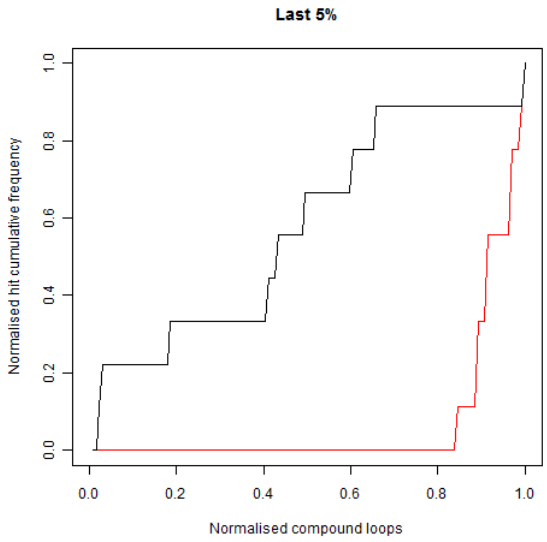


Figure 5.3: Examples of rare category compound curves

The Active Learning curves and deficiency curves were built using the full Maybridge Hitfinder library; all simulations are displayed in Appendix B. The following examples use the data for TcDHFR as the target and PfRdhfr as the second parasite; they show typical curve characteristics, and the features related to the algorithm under test are also described:

Active k-optimisation, active learning curves

Red: active vs target
Blue: active vs Hs
Green: active vs other parasite
Orange: toxic

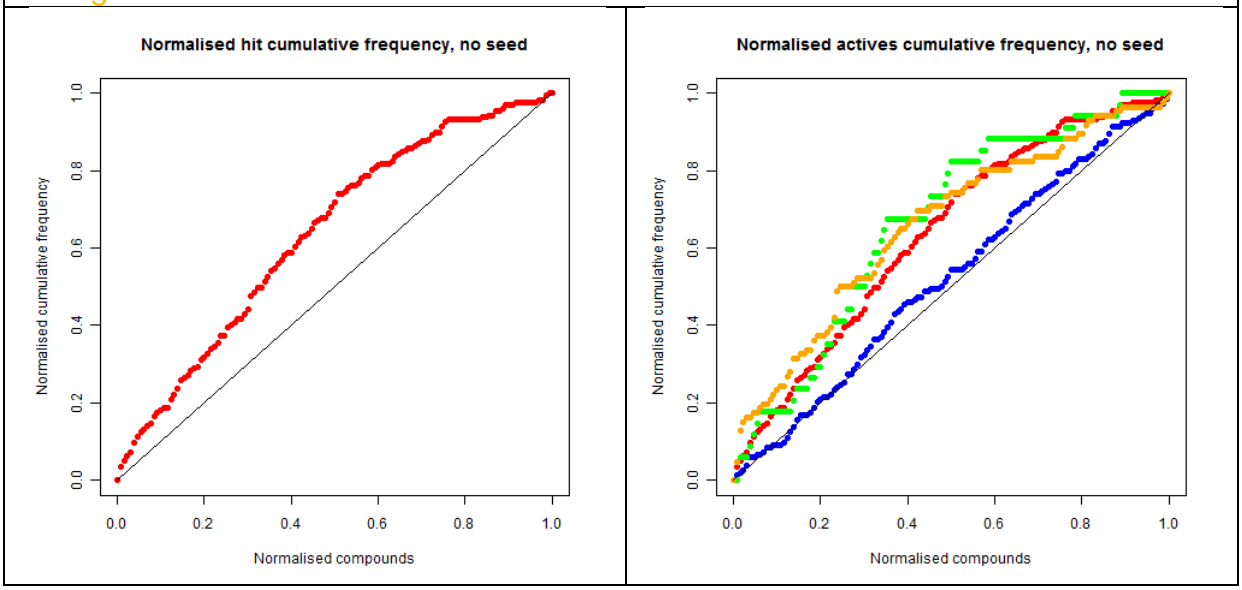


Figure 5.4: Examples of *Active k-optimisation* learning curves

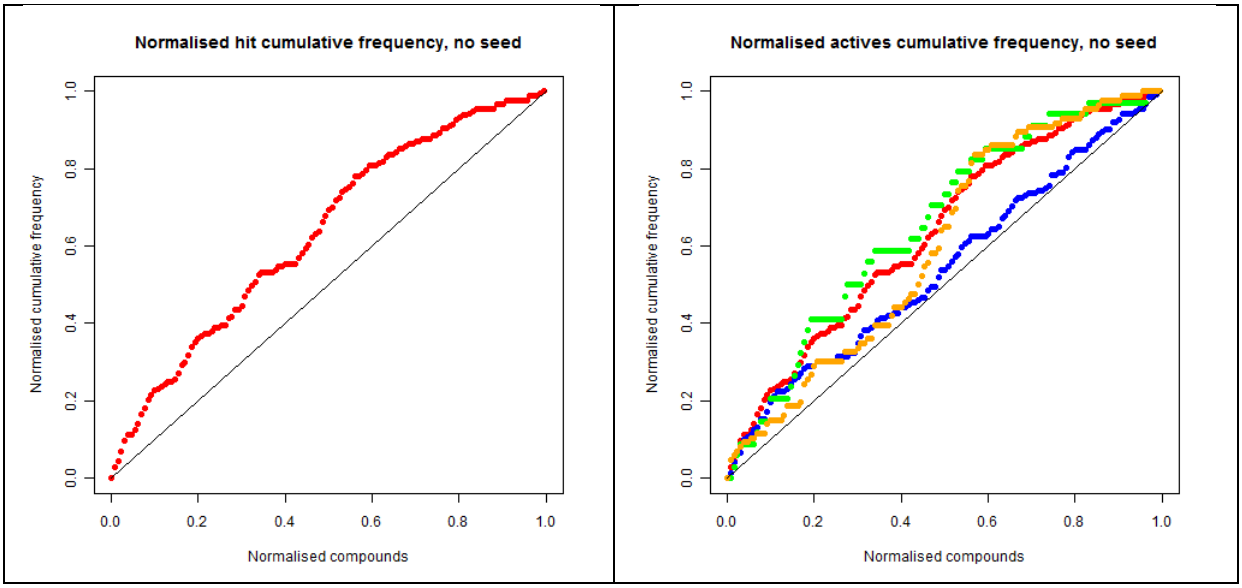


Figure 5.5: Examples of *SimplyGreedy* learning curves

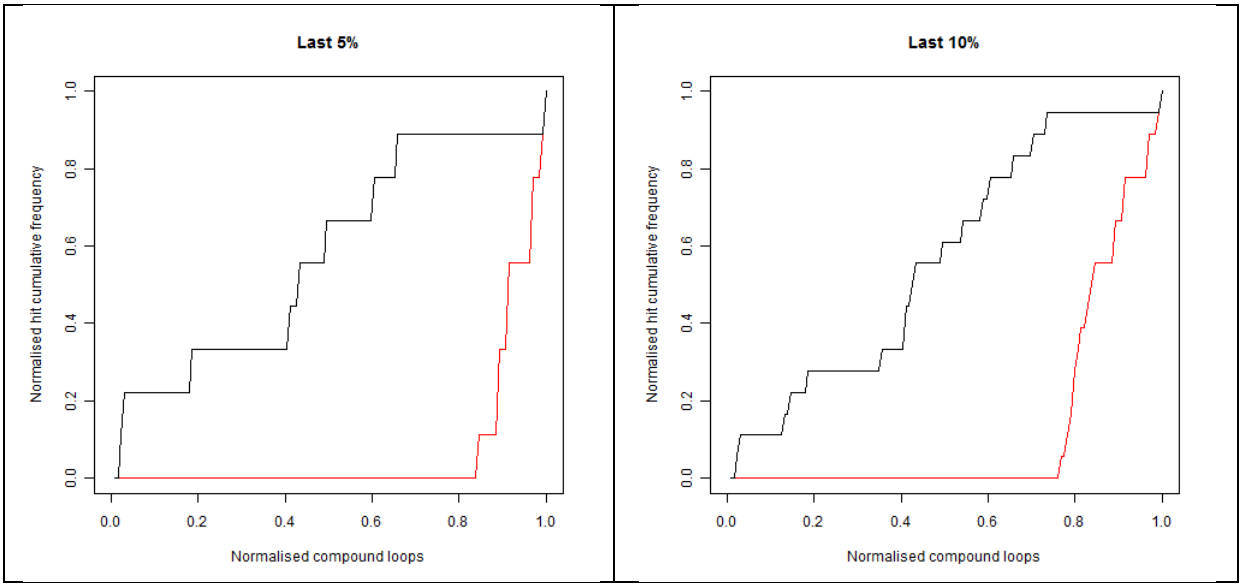


Figure 5.6: Rare category deficiency curves, last 5% & 10% actives: *active k-optimisation* (black) versus *SimplyGreedy* (red)

In general, *active k-optimisation* and *SimplyGreedy* AL curves are fairly similar when using common seed compound sets. The benefit of the former's strategy for exploration of the chemical space becomes apparent when the rare category compounds are identified and examined: these are typically found at a steady rate throughout this simulation.

Preclustering, $TS > 0.40$

Red: active vs target
Blue: active vs Hs
Green: active vs other parasite
Orange: toxic

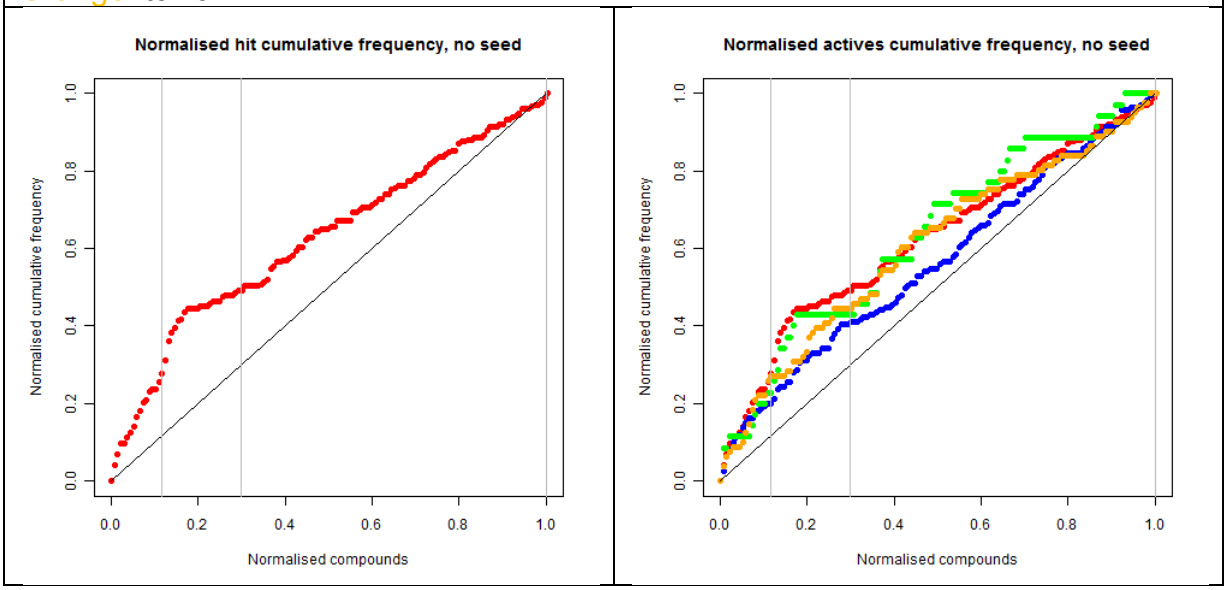


Figure 5.7: Examples of *Preclustering* learning curves, TS > 0.40

Preclustering, $TS > 0.45$

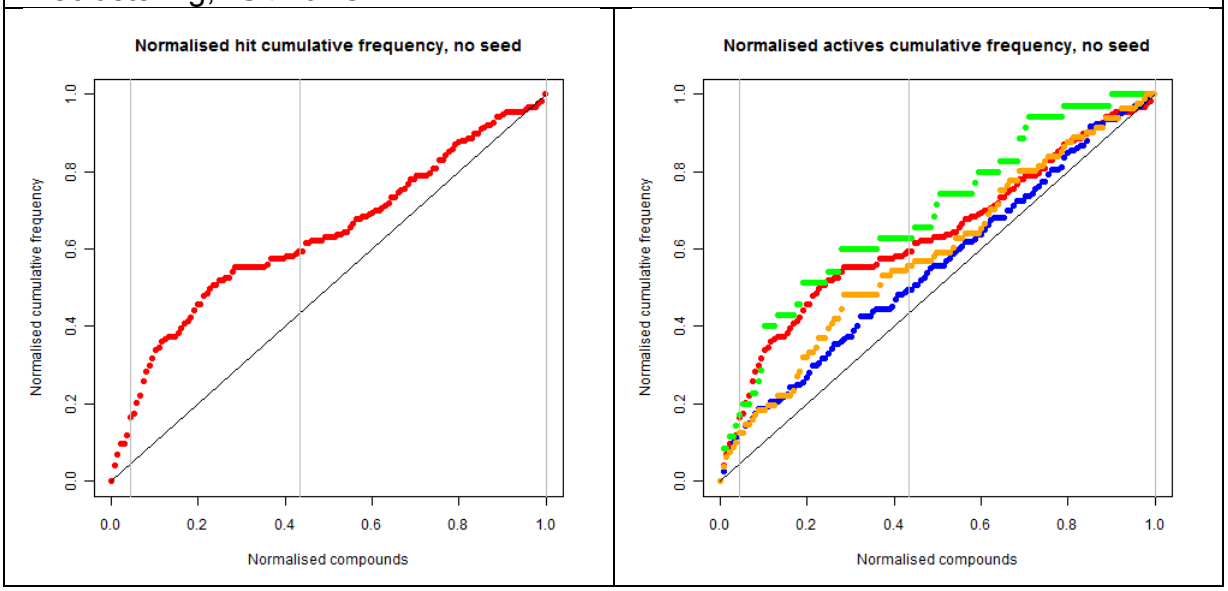


Figure 5.8: Examples of *Preclustering* learning curves, TS > 0.45

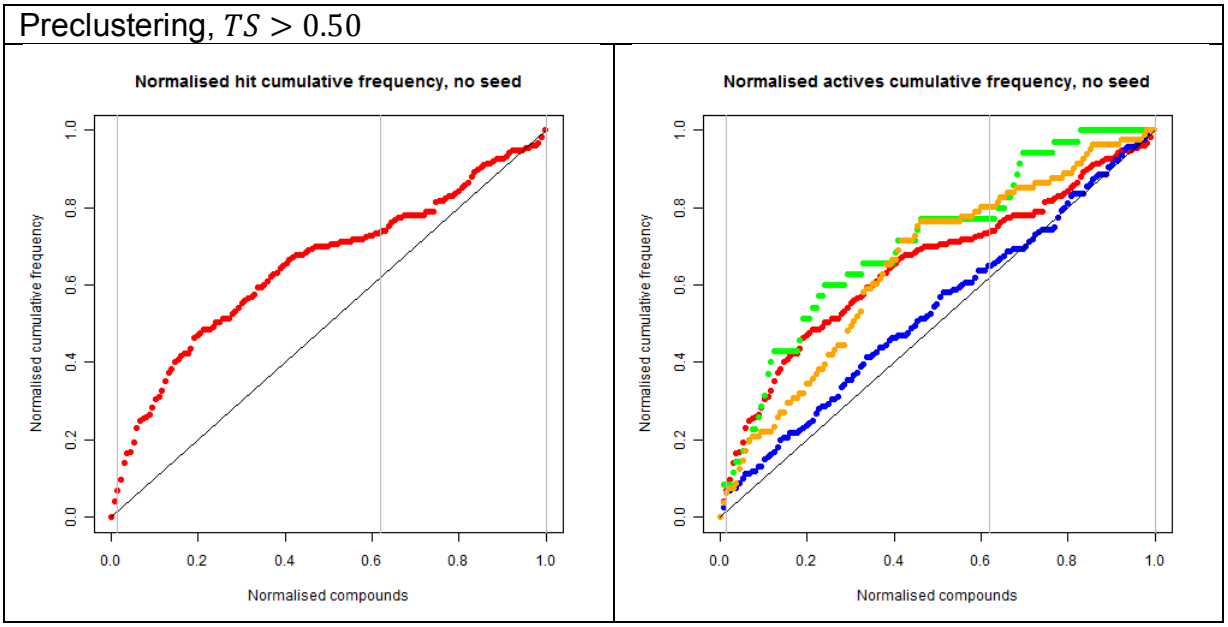


Figure 5.9: Examples of *Preclustering* learning curves, TS > 0.50

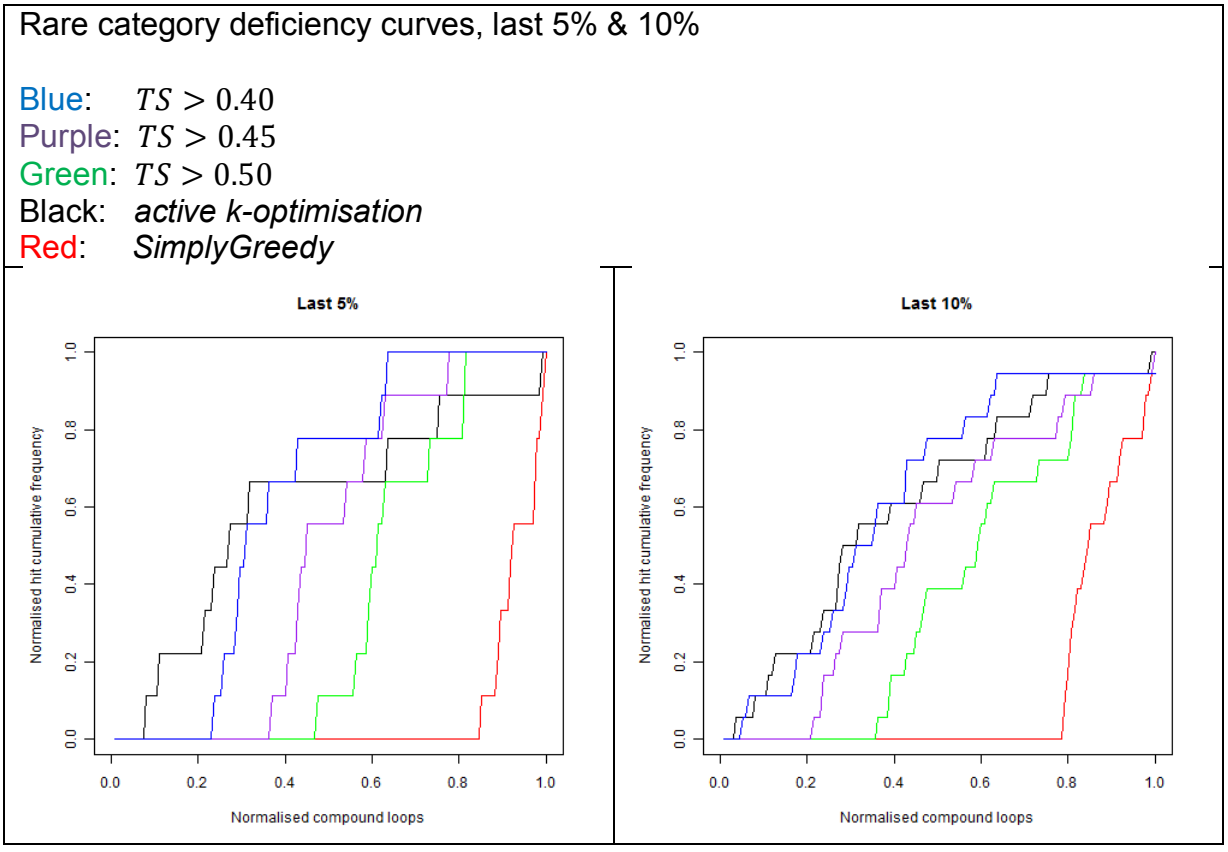


Figure 5.10: Rare category deficiency curves, last 5% & 10% actives: *active k-optimisation* (black) versus *SimplyGreedy* (red) versus preclustering at three TS limits

For the preclustering algorithm, the effect of different limits for $TS > x$ can readily be seen in the three phases of candidate compounds:

- Phase 1: testing those unknowns that are similar to seed active compounds.
- Phase 2: greedy selection from unknowns which are not similar to inactive compounds.
- Phase 3: the remaining unknowns are ranked in ascending order of number of similar inactive compounds.

High limits for TS result in a short initial phase, with a commensurately long second phase. As the threshold for TS falls, the first two phases expand and contract accordingly.

The exploration of Phase 3 could be extended to incorporate selections based on similarity to active seeds. The existing method for this tail section doesn't seem to offer much performance benefit over a random selection strategy. This suggestion will be included in options for further work.

Transfer Learning

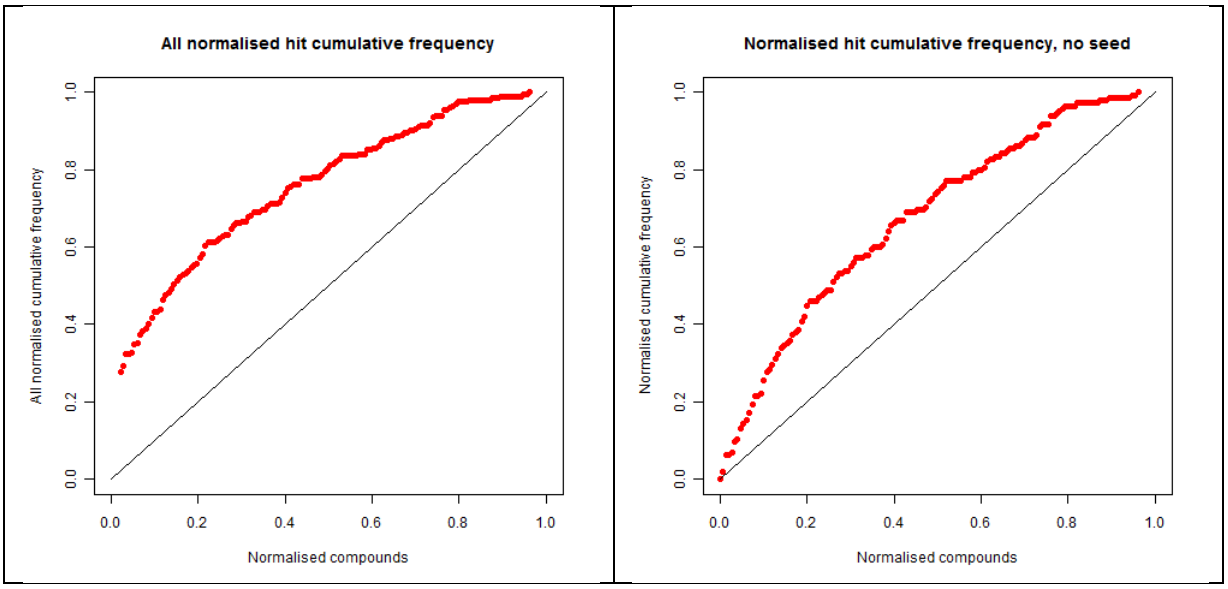


Figure 5.11: Examples of *TranferLearning* learning curves

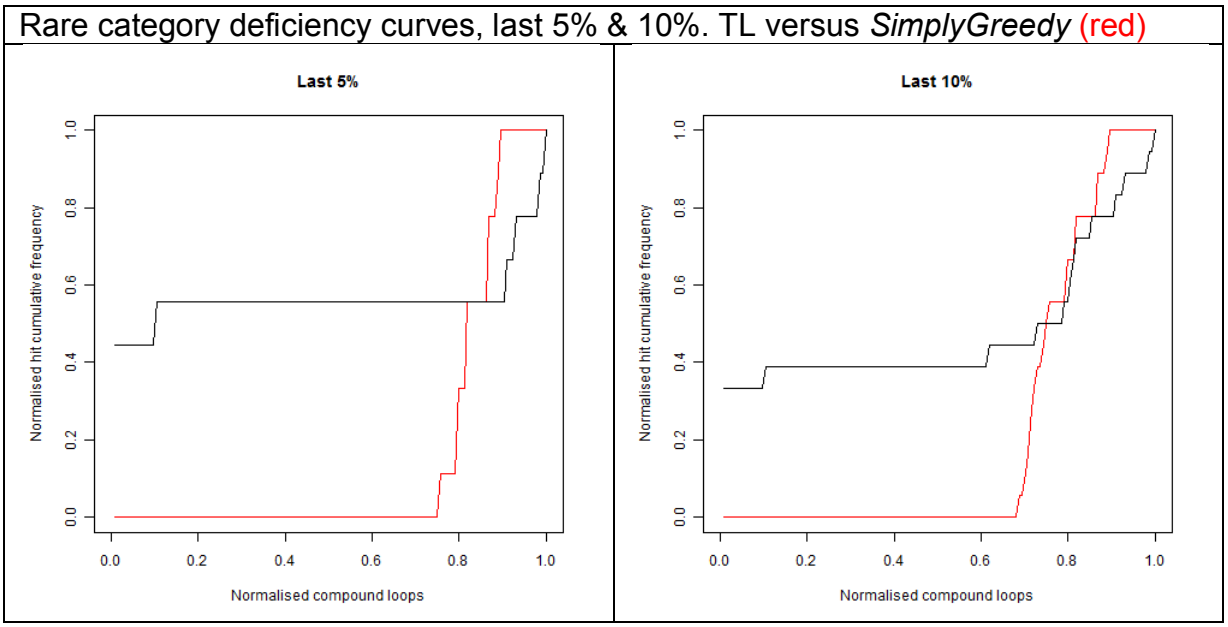


Figure 5.12: Rare category deficiency curves, last 5% & 10% actives: *TransferLearning* (black) versus *SimplyGreedy* (red)

The preceding TL cumulative frequency curves show the strong boost provided by the exogenous active seeds at the start of the simulation. The right-hand curve shows how the selection progresses if the initial active and inactive seed results are stripped from the dataset. In the endogenous learning curves, the proportion of

active:inactive will typically be the same as the background level for the full library, whereas the proportion of active seeds in transfer learning regimes could be much higher if there is strong similarity between targets.

Similarly, rare category detection is given an early boost by having some compounds from the rare chemical space in the seed group. The greedy selection method used by the TL algorithm means that the remaining proportion of rare category compounds are not likely to be found until the late stages of the simulation.

Transfer Learning with preclustering

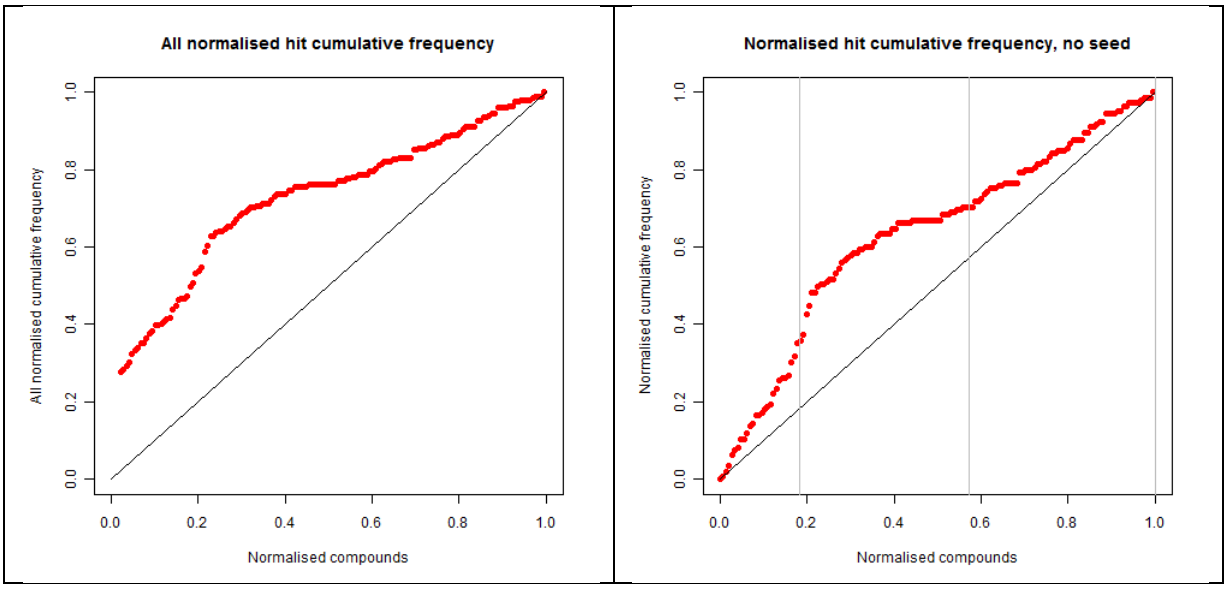


Figure 5.13: Examples of *TransferLearning with preclustering* learning curves

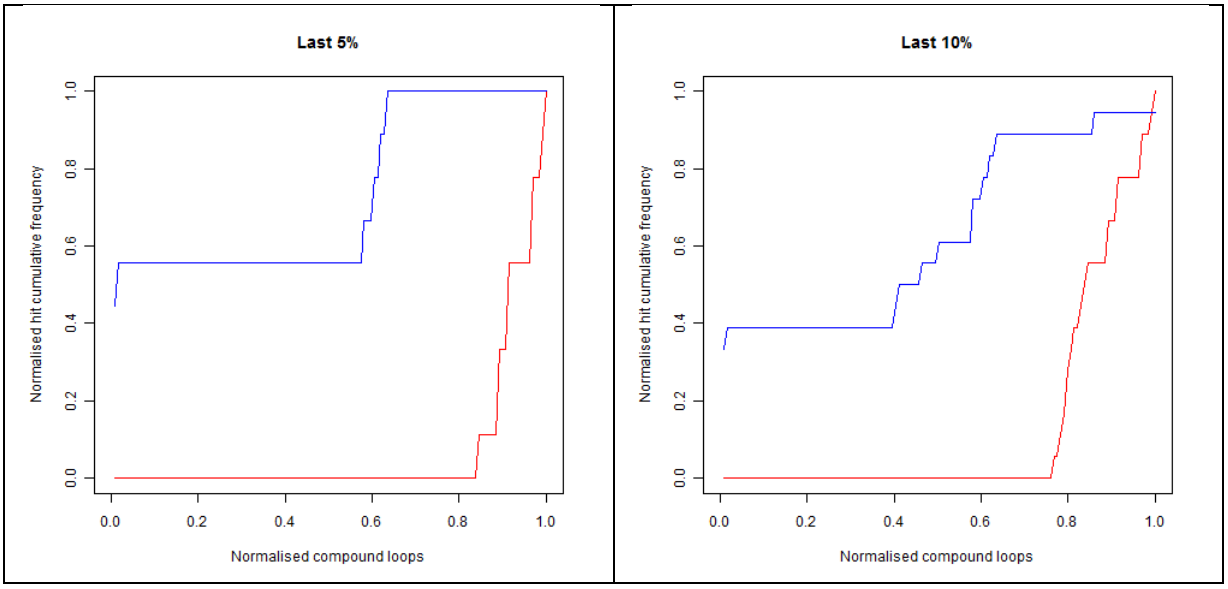


Figure 5.14: Rare category deficiency curves, last 5% & 10% actives: *TransferLearning with preclustering* (black) versus *SimplyGreedy* (red)

Combining the Transfer Learning and preclustering ideas, another boost is seen for the detection of rare category compounds. The early, initial seed rarities are augmented by those in the inactive-dissimilar compounds in Phase 2.

This combination gives the strongest overall performance of all the algorithms tested *in silico*, and has the advantage of being based on classification data which reduces computational complexity.

5.2 Results for endogenous simulations

Each AL algorithm was evaluated extensively using multiple parasite-protein datasets. The algorithms that operated on a single assay were tested using several different sets of plates to provide the seed data. The algorithms that included transfer learning elements (see section 5.3) were evaluated by building learning curves for each parasite-protein target when seeded using different sets of known parasite-active compounds.

Parasite	Protein - screen no.	Completed simulation iterations				
		Active k opti'mstn	*Simply Greedy*	*Preclustering*, $TS > x$		
				$x = 0.40$	$x = 0.45$	$x = 0.50$
Pv	DHFR-TS3	3	3	3	3	3
PfR		3	3	1	1	1
Tb	DHFR-TS4	3	3	-	-	-
Sm		3	3	-	-	-
PfR	DHFR-TS5	3	3	2	2	2
Tc		3	3	1	1	1
Pv	DHFR-TS6	11	11	10	10	10
Pf		3	3	3	3	3
Lm	DHFR-TS7	3	3	-	-	-
PvR		3	3	2	2	2
Tb	NMT-1	3	3	1	1	1
Pv		3	3	2	2	2
Sm	NMT-2	-	-	-	-	-
Tc		8	9	4	4	4
Sm	PGK-1	2	2	-	-	-
Tc		3	3	1	1	1
Tb	PGK-2	2	2	-	-	-
Pv		3	3	-	-	-

Table 5.1: Experiments using endogenous single mass screen datasets

At least three iterations of each algorithm were attempted for all parasite-protein combinations; more were started for PvDHFR-TS3 (11) and TcNMT (9). Not all *active k-optimisation* and *SimplyGreedy* simulations ran to completion, due to the level of active compounds being too low to support continuation.

All learning curves are reported in Appendix B, together with deficiency measurements for all experiments. The tables in this section contain the mean of the deficiencies for all parasite-protein combinations, including those for the rare category compounds. Experiment means marked * are based on less than three

curves. Two parasite-protein combinations (PvDHFR & TcNMT) were run over a larger number of iterations to examine the repeatability of the simulations.

Figures 5.15 to 5.17 compare the performance of *SimplyGreedy, active k-optimisation* and *preclustering* algorithms, by using boxplots built from full data sets (i.e. where an individual seed/unknown data set has been used in a simulation for each strategy). Whilst this approach will not use all of the available experiments in Table 5.1, it will allow a direct comparison of the performance of these algorithms.

Parasite	Protein – screen no.	Completed simulation iterations				
		Active k-opti'mstn	*Simply Greedy*	*Preclustering*, $TS > x$		
				$x = 0.40$	$x = 0.45$	$x = 0.50$
Pv	DHFR-TS3	0.64	0.67	0.72	0.76	0.74
PfR		0.91	0.88	0.91*	0.94*	0.88*
Tb	DHFR-TS4	0.69	0.63	-	-	-
Sm		0.84	0.79	-	-	-
PfR	DHFR-TS5	1.01	0.88	0.84*	0.84*	0.87*
Tc		0.73	0.72	0.78*	0.76*	0.77*
Pv, mean	DHFR-TS6	0.62	0.67	0.78	0.79	0.77
Pv, SD		0.012	0.016	0.042	0.025	0.039
Pf		0.83	0.74	0.77	0.80	0.76
Lm	DHFR-TS7	0.83	0.66	-	-	-
PvR		0.83	0.70	0.76*	0.78*	0.77*
Tb	NMT-1	0.72	0.78	0.92*	0.86*	0.84*
Pv		0.83	0.79	0.82*	0.80*	0.77*
Sm	NMT-2	-	-	-	-	-
Tc, mean		0.80	0.80	0.91	0.87	0.83
Tc, SD		0.018	0.015	0.023	0.018	0.041
Sm	PGK-1	0.71*	0.62*	-	-	-
Tc		0.93	0.86	-	-	-
Tb	PGK-2	0.79*	0.53*	-	-	-
Pv		0.95	0.82	-	-	-

Table 5.2: Mean deficiencies for single mass screen experiments

Parasite	Protein - screen no.	Completed simulation iterations			
		Active k-opti'mstn	*Preclustering*, $TS > x$		
			$x = 0.40$	$x = 0.45$	$x = 0.50$
Pv	DHFR-TS3	0.51	0.40	0.41	0.59
PfR		0.40	0.52*	0.44*	0.51*
Tb	DHFR-TS4	0.41	-	-	-
Sm		0.37	-	-	-
PfR	DHFR-TS5	0.55	0.40	0.44	0.58
Tc		0.44	0.42*	0.52*	0.62*
Pv, mean	DHFR-TS6	0.45	0.50	0.50	0.65
Pv, SD					
Pf		0.42	0.37	0.42	0.54
Lm	DHFR-TS7	0.62	-	-	-
PvR		0.57	0.38*	0.46*	0.63*
Tb	NMT-1	0.43	0.44*	0.56*	0.68*
Pv		0.42	0.37*	0.38*	0.50*
Sm	NMT-2	-	-	-	-
Tc, mean		0.40	0.45	0.52	0.69
Tc, SD		0.052	0.057	0.056	0.051
Sm**	PGK-1	-	-	-	-
Tc		0.39	0.47*	0.51*	0.69*
Tb	PGK-2	-	-	-	-
Pv		0.71	-	-	-

Table 5.3: Rare category deficiencies (last 5%)

Parasite	Protein/ screen	Completed simulation iterations			
		Active k-opti'mstn	*Preclustering*, $TS > x$		
			$x = 0.40$	$x = 0.45$	$x = 0.50$
Pv	DHFR-TS3	0.47	0.42	0.42	0.58
PfR		0.48	0.49*	0.42*	0.50*
Tb	DHFR-TS4	0.23	-	-	-
Sm		0.47	-	-	-
PfR	DHFR-TS5	0.50	0.45	0.44	0.57
Tc		0.48	0.47*	0.54*	0.64*
Pv, mean	DHFR-TS6	0.40	0.52	0.52	0.65
Pv, SD		0.047	0.038	0.053	0.062
Pf		0.44	0.44	0.46	0.57
Lm	DHFR-TS7	0.54	-	-	-
PvR		0.52	0.49*	0.51*	0.66*
Tb	NMT-1	0.38	0.51*	0.58*	0.71*
Pv		0.42	0.45*	0.42*	0.50*
Sm	NMT-2	-	-	-	-
Tc, mean		0.38	0.46	0.50	0.67
Tc, SD		0.030	0.036	0.046	0.076
Sm	PGK-1	-	-	-	-
Tc		0.45	0.48*	0.53*	0.71*
Tb	PGK-2	-	-	-	-
Pv		0.70	-	-	-

Table 5.4: Rare category deficiencies (last 10%)

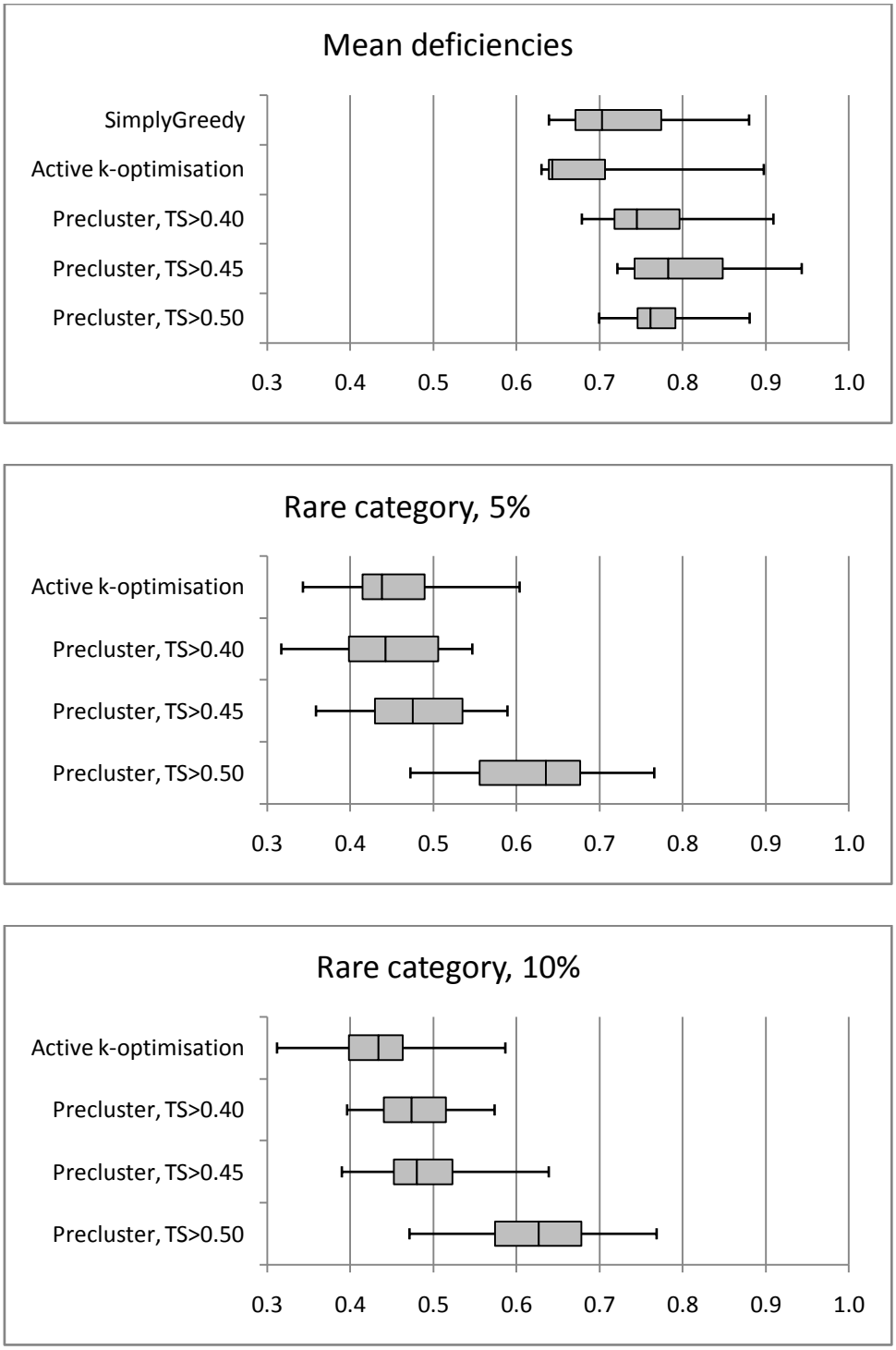


Figure 5.15: Boxplots for collated deficiency measurements: mean and rare category compounds (5% & 10%). A comparison of the *SimplyGreedy*, *active k-optimisation* and *preclustering* algorithms, using simulation data where all strategies were tested.

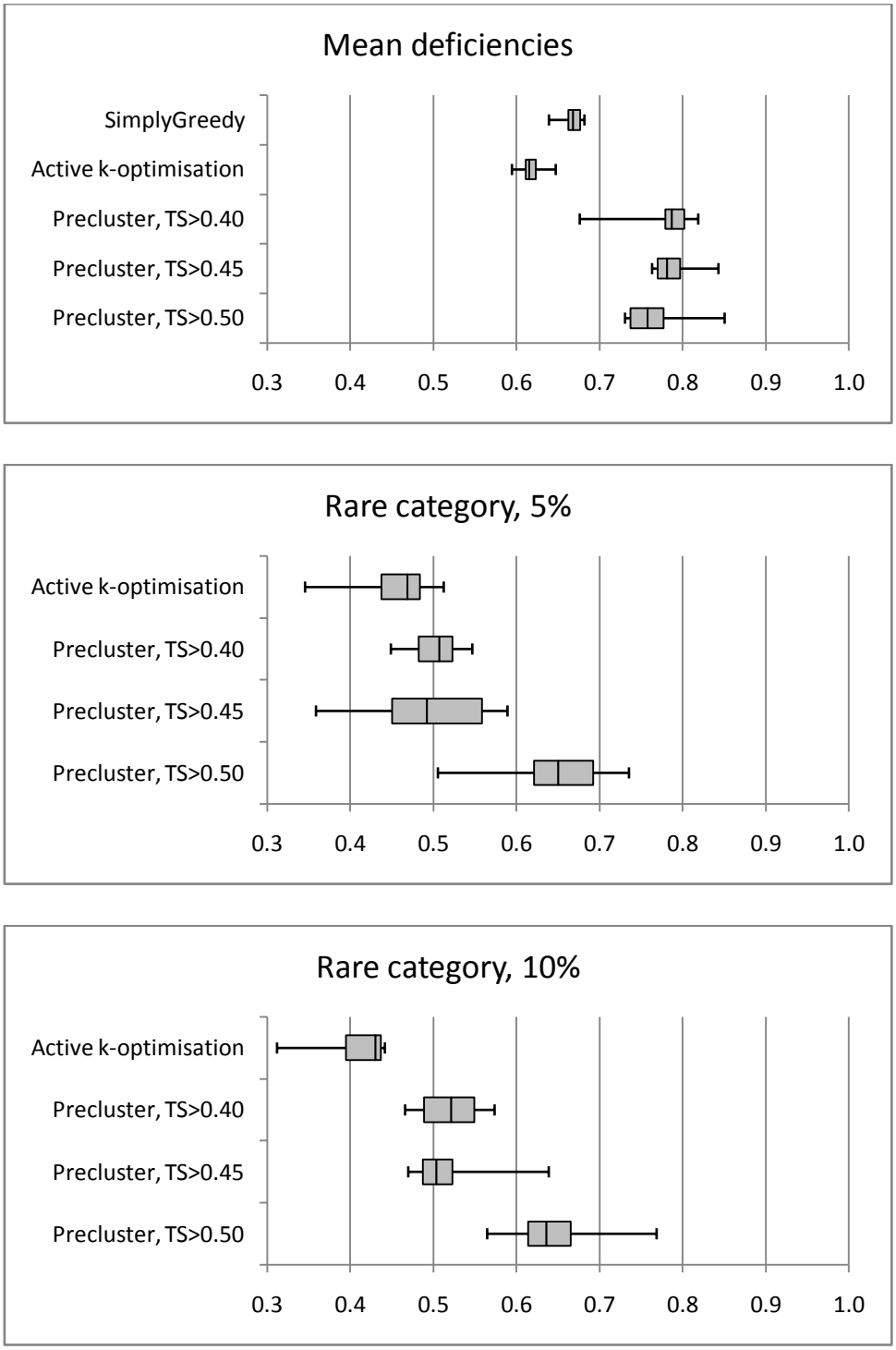


Figure 5.16: Boxplots for deficiency measurements in 10 PvDHFR TS6 simulations: mean and rare category compounds (5% & 10%). A comparison of the *SimplyGreedy, active k-optimisation* and *preclustering* algorithms, using simulation data where all strategies were tested.

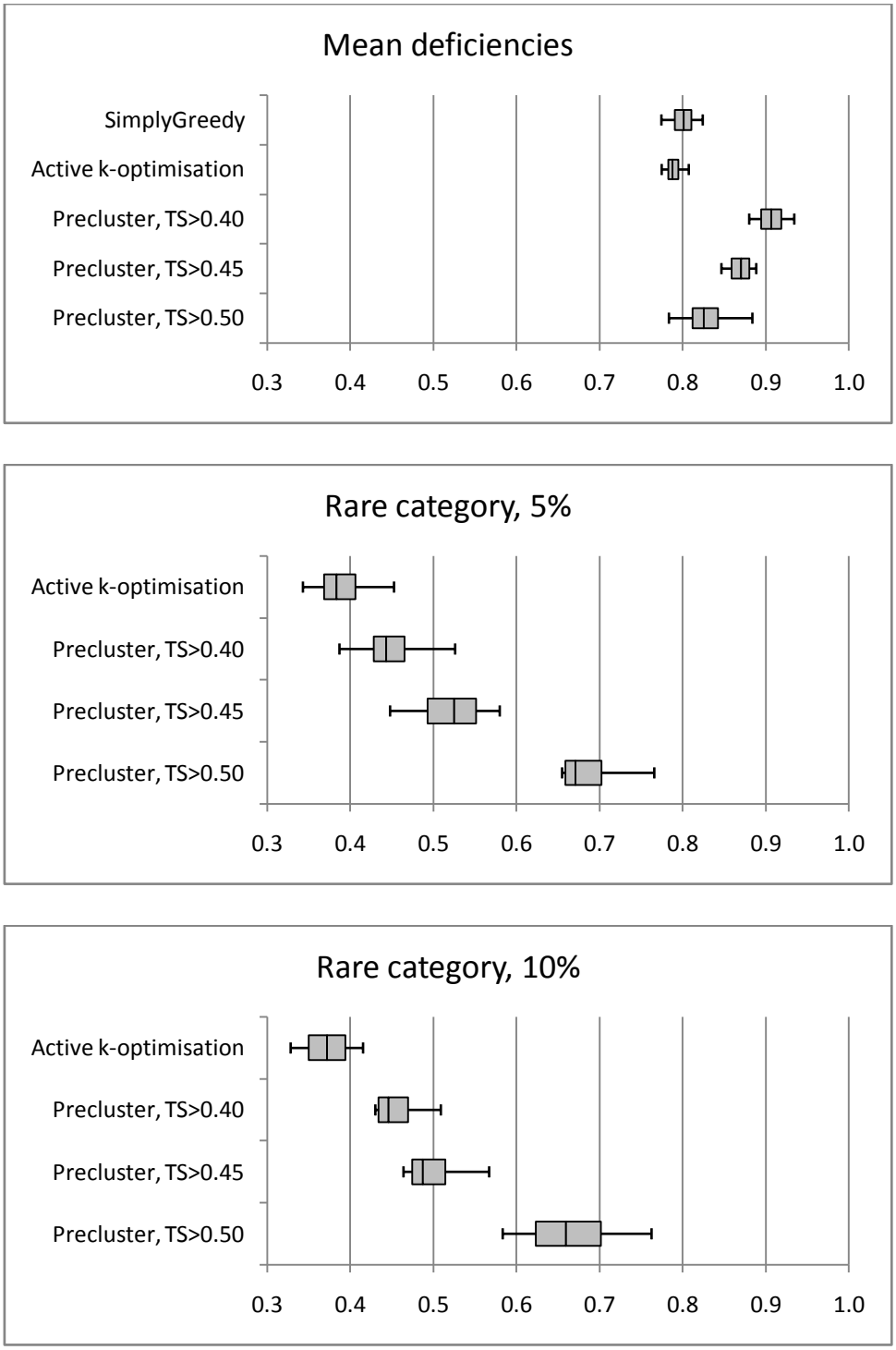


Figure 5.17: Boxplots for deficiency measurements in 4 TcNMT simulations: mean and rare category compounds (5% & 10%). A comparison of the *SimplyGreedy*, *active k-optimisation* and *preclustering* algorithms, using simulation data where all strategies were tested.

5.3 Results for Transfer Learning simulations

5.3.1 Transfer Learning

In its simplest form, the Transfer Learning simulation could only be run once for each set of seed compounds. However, each target could be seeded with one of several sets of transfer seeds, so multiple applications and resultant cross-strain evaluations were possible.

The initial work (see Tables 5.5 to 5.7) concentrated on identifying the beneficial effect of using lists of active compounds from previous screens; these simulations were built using a greedy search after taking the initial seed compounds (the TL training set) as a proxy partial mass screen. For each parasite-protein combination, a simulation was also run using a HsDHFR dataset; it was expected that this might identify interference due to general toxicity against the target (possibly a significant effect for targets with low numbers of true active candidates).

Parasite	Protein/ screen	DHFR seeds					NMT seeds		
		Hs	Pf	$Pv_{(TS6)}$	Tb	Tc	Pv	Tb	Tc
Pv	DHFR-TS3	0.68	0.59	**0.29**	0.58	**0.51**	0.60	**0.50**	0.59
PfR		**0.35**	0.65	0.71	0.88	0.66	0.74	0.74	0.75
Tb	DHFR-TS4	0.57	0.66	**0.43**		**0.45**	0.60	**0.51**	0.57
Tc	DHFR-TS5	0.55	**0.43**	**0.47**	0.66			0.54	**0.50**
Pv	DHFR-TS6	0.66	**0.49**		**0.59**	**0.55**	**0.58**	**0.51**	**0.59**
Pf		0.54		0.45		**0.43**	0.55	0.57	0.54
PvR	DHFR-TS7	0.63	**0.52**	**0.42**	0.65	**0.50**	**0.57**	**0.52**	**0.56**
Tb	NMT-1	0.74	0.66	**0.59**	0.72	0.66	**0.58**		0.67
Pv	NMT-1	0.74	0.62	**0.63**	0.77	0.65		0.53	0.50
Tc	NMT-2	0.56	0.62	0.61	0.76	0.60	**0.49**	0.65	
Mean deficiency, all targets		0.60	0.58	0.51	0.70	0.56	0.59	0.56	0.59
Mean of deficiencies: 0.59 *Mean of core PfDHFR/PvDHFR seed deficiencies: 0.60									

Note *: the experiments in the nine cells highlighted in green were run in all TL simulations, and could therefore be used as a core comparison between the simple and TL+preclustering methods. Results in bold text indicate those which were significantly better than when compounds active versus HsDHFR were used as the TL seed.

Table 5.5: Deficiencies for TL experiments, single seed

Parasite	Protein/ screen	DHFR seeds					NMT seeds		
		Hs	Pf	$Pv_{(TS6)}$	Tb	Tc	Pv	Tb	Tc
Pv	DHFR-TS3	0.98	1.06	**0.67**	1.07	**0.76**	0.99	**0.68**	0.91
PfR		**0.36**	0.90	0.89	1.15	0.97	0.98	1.04	0.89
Tb	DHFR-TS4	0.83	1.23	1.22		0.83	1.23	0.80	0.83
Tc	DHFR-TS5	1.02	**0.64**	**0.52**	1.15			**0.76**	0.90
Pv	DHFR-TS6	1.06	**0.75**		1.07	**0.75**	**0.83**	0.90	0.91
Pf		0.88		**0.36**		**0.61**	0.74	0.74	**0.62**
PvR	DHFR-TS7	0.95	0.79	**0.61**	0.89	**0.61**	**0.70**	**0.69**	1.05
Tb	NMT-1	1.12	1.04	0.95	1.12	0.88	**0.67**		1.11
Pv	NMT-1	1.05	1.14	0.94	1.15	1.05		**0.85**	0.95
Tc	NMT-2	**0.76**	1.08	0.97	0.98	1.08	**0.54**	0.97	
Mean deficiency, all targets		0.90	0.96	0.79	1.07	0.84	0.84	0.83	0.91
Mean of deficiencies: 0.89 Mean of core PfDHFR/PvDHFR seed deficiencies: 0.97									

Table 5.6: Rare category deficiencies (last 5%), single seed

Parasite	Protein/ screen	DHFR seeds					NMT seeds		
		Hs	Pf	$Pv_{(TS6)}$	Tb	Tc	Pv	Tb	Tc
Pv	DHFR-TS3	1.01	1.02	**0.56**	1.00	**0.84**	1.02	**0.76**	0.92
PfR		**0.34**	0.92	0.93	1.11	0.96	1.02	1.01	0.96
Tb	DHFR-TS4	1.01	1.30	1.28		**0.45**	1.03	1.00	1.02
Tc	DHFR-TS5	0.83	**0.58**	**0.67**	1.15			0.80	0.76
Pv	DHFR-TS6	1.06	0.90		0.89	0.90	**0.84**	0.87	0.87
Pf		0.86		**0.57**		0.70	**0.67**	0.72	0.72
PvR	DHFR-TS7	0.93	**0.71**	**0.69**	0.97	**0.72**	**0.72**	**0.61**	0.95
Tb	NMT-1	1.12	0.92	0.84	1.01	0.88	**0.80**		0.96
Pv	NMT-1	1.05	1.09	1.03	1.15	0.94		**0.83**	**0.83**
Tc	NMT-2	**0.73**	0.90	0.78	1.03	0.73	**0.57**	0.93	
Mean deficiency, all targets		0.89	0.93	0.82	1.04	0.79	0.83	0.84	0.89
Mean of deficiencies: 0.88 Mean of core PfDHFR/PvDHFR seed deficiencies: 0.92									

Table 5.7: Rare category deficiencies (last 10%), single seed

5.3.2 Transfer Learning with preclustering

Layering the preclustering and TL algorithms allowed a further round of simulations to be conducted. In addition, the preclustering Tanimoto Similarity threshold could be adjusted to observe any effects on how the unknown compounds were categorised versus the active and inactive seeds.

Two rounds of experiments were conducted. The parameters were chosen based on observations from the earlier preclustering work (section 5.2). The first round used a similarity threshold of $TS > 0.60$ to identify the initial candidate list, followed by a threshold of $TS > 0.40$ for the body of the experiments. This was designed to restrict the number of seed compounds, just in case this group became too large with a subsequent detrimental effect on inactive-dissimilar unknown compounds.

The second round used $TS > 0.40$ as the threshold throughout the experiments; this brought the experiments closer to those reported in section 5.2. These experiments were only run using PvDHFR or PfDHFR as the transfer seeds. A summary of the deficiency measurements for this work is given in Table 5.8, together with the simple TL results and the ranges of results from earlier endogenous work.

Row ID	Description		Mean, all	Mean, core
a	Transfer Learning	Main AL curve	0.59	0.60
b		Rare category, 5%	0.89	0.97
c		Rare category, 10%	0.88	0.92
d	TL + precluster, version 1 (0.6 then 0.4)	Main AL curve	0.70	0.73
e		Rare category, 5%	0.54	0.55
f		Rare category, 10%	0.52	0.53
g	TL + precluster, version 2 (all 0.4)	Main AL curve	0.65	0.69
h		Rare category, 5%	0.52	0.58
i		Rare category, 10%	0.51	0.55
	SimplyGreedy (range for equivalent simulations)	Main AL curve	0.63 - 0.88	
	Preclustering (range)	Main AL curve	0.72 - 0.92	
		Rare category, 5%	0.37 - 0.69	
		Rare category, 10%	0.42 - 0.71	
	Active k-optimisation (range)	Main AL curve	0.62 - 0.91	
		Rare category, 5%	0.40 - 0.57	
		Rare category, 10%	0.40 - 0.52	

Table 5.8: Mean deficiencies for TL simulations

The main curve deficiencies for the two TS threshold variants were not significantly different from each other, and were in line with the better results from endogenous simulations using *SimplyGreedy* and *active k-optimisation*. Both TL+precluster variants were were appreciably better at rare category detection than the simple TL method, and were both in line with typical results from *active k-optimisation* (note: a direct comparison could not be made between endogenous and exogenous results as, by definition, the simulation used different seed/unknown compounds sets).

The differences in deficiency between simple TL and TL+preclustering given in Table 5.9 confirm the improvement in rare category detection with the latter method, albeit with a slight reduction in overall selection efficiency.

Description		Calculation (from row data in Table 5.8)	Mean, all	Mean, core
Version 1 versus simple TL	Main AL curve	a – d	-0.11	-0.13
	Rare category, 5%	b – e	0.40	0.42
	Rare category, 10%	c – f	0.38	0.39
Version 2 versus simple TL	Main AL curve	a – g	-0.11	-0.10
	Rare category, 5%	b – h	0.37	0.39
	Rare category, 10%	c – i	0.34	0.37

Table 5.9: Deficiency differences, TL simulations

The full result tables for this work are given in Appendix B.5. These tables are split into two variants, and an overall evaluation of the algorithm's performance versus simple Transfer Learning:

- Tables B.5 (1 to 3) report results where preclustering Phase 1 was conducted using a threshold of $TS > 0.60$ with Phase 2 using a threshold of $TS > 0.40$. This was designed to build on the potentially strong seed dataset; this should have the effect of creating a longer initial phase than the simple preclustering routines in section 5.2. It was also expected to limit the number of unknown compounds that were labelled as similar to inactive seeds.

- Tables B.5 (4 to 6) were built used a preclustering routine where all phases used the threshold $TS > 0.40$. Simulations were only run using two of the exogenous seed datasets; this was to examine whether the conditions applied to the immediately preceding experiments were too conservative.

- Combining the TL and preclustering mechanisms was expected to deliver better overall performance for AL than the baseline *SimplyGreedy* selections. It was also expected to boost identification of rare category compounds at an earlier stage in the process. The difference in deficiencies for the main TL/preclustering learning curves compared to using TL only are shown in tables B.5 (7 to 12).

5.4 Relative speed and complexity of algorithms

The time taken for the selection process for each loop of *active k-optimisation* was dominated by generation and evaluation of the seed matrix. After each loop, multiple data points for each newly analysed compound were added to the seed in preparation for calculating the next set of unknowns to be tested.

In contrast to the selection processes based on categorised seed compounds (i.e. all processes other than *active k-optimisation*) the confirmation results were used to merely make a single categorisation of each compound. In all phases of these processes, once potential candidates had been ranked in terms of their similarity to the initial seed compounds, any subsequent re-ranking was based entirely on analysis of any freshly identified active compounds. Both of these simplifications led to a much shortened selection process.

During early work with *active k-optimisation*, benchmarking between my PC and the Leuven computing facility was conducted. Partial simulations of 30 loops length were conducted and timed, and the results are given in Table 5.10.

Loop	PC simulation	PINAC19 Simulation 1	PINAC19 Simulation 2
Initial	85’24	82’12	82’44
2nd	27,01	24’02	24’12
3rd	28’15	24’55	25’05
4th	29’03	25’51	26’00
5th	30’00	26’46	26’57
28th	54’36	48’19	48’26
29th	55’44	49’10	49’20
30th	56’53	50’07	50’09

Table 5.10: Timing comparisons for *active k-optimisation* simulations

Full simulations of >100 loops were found to take in excess of 1 week to complete; this is in contrast to the selection processes based on simpler categorised seed compounds where, for example, a 148 loop transfer learning with preclustering simulation took 12½ hours to complete (this will have been close to the longest duration of any of the non- *active k-optimisation* simulations).

The large difference between these approaches isn't necessarily a barrier to using the more complex method, as more computing power could be used to accomplish the required tasks in an appropriate time. However, when using much larger chemical libraries it is clear that selection times will be stretched further, and if simpler algorithms also give suitable outputs then they will be seen as more competitive. It is also evident that the matrix computations required by the *active k-optimisation* algorithm are memory intensive: another potential limitation.

Another factor to consider would be the time taken to physically conduct each experiment loop. Eve's assay is comparatively long at 40 hours, but it is known that pharmaceutical companies strive for short duration assays where possible, and for an Active Learning approach to be useful it would need to generate selections that do not substantially lag the current round of experiments.

5.5 Discussion

Comparisons across all endogenous algorithm variants (Tables 5.2 to 5.4, Figures 5.15 to 5.17)

By taking the mean of all comparable deficiencies for the endogenous algorithms, it was possible to compare their performance when selecting both active compounds and rare category compounds. The full AL curves for *SimplyGreedy* and *active k-optimisation* had marginally better deficiency results than when *preclustering*.

In contrast, the *SimplyGreedy* algorithm is (by definition) a poor selector of rare category compounds, and comparison made at the 5% and 10% levels shows that the *active k-optimisation* performs well in this respect, due to its chemical space exploration strategy. The three variants of the *preclustering* algorithm ($TS > [0.40,0.45,0.50]$) are also strong performers, but reduce in effectiveness with increasing TS limits; it has already been suggested that this relates to the increasing number (with an increasing TS limit) of unlabelled unknowns for evaluation in Phase 2 of this algorithm.

Comparisons between exogenous algorithms (Tables 5.8 & 5.9)

When an exogenous set of parasite-active compounds is applied as a *TL* seed, a strong improvement is seen in the rate of selection of active candidates when compared to all the endogenous AL curves. Identification of rare category compounds shows a marginal improvement over *SimplyGreedy*.

Using *preclustering* together with *TL* reduces the overall effectiveness slightly, but provides a large boost to rare category detection. The combination of the different attributes of these methods results in a performance trade off from both sides, but the overall effect is very encouraging. Slightly different strategies for Phase 1 of this process were compared; there seemed to be a marginal improvement when *preclustering* at $TS > 0.40$ was compared to $TS > 0.60$, but this effect was much lower than seen in the endogenous *preclustering* scoping work.

It should be noted that the above comments relate to the generalised *TL* processes. Depending on the choice of exogenous seed, much larger gains in active compound and rare category compound deficiencies were observed (e.g. PvDHFR as a seed

for TcDHFR). This suggests that much better gains can be made by tailoring the seed group, depending on knowledge or prediction of cross-strain target similarity. When comparing the best of the *TL/preclustering* simulations with those of the *active k-optimisation* process, it is clear that the gains offered by the former are potentially much larger.

Strain-by-strain analysis of Transfer Learning versus endogenous Active Learning algorithms (**Tables 5.5 to 5.7)**

The simple *TL* simulations used several active seed data sets from the DHFR and NMT mass screens, including those of HsDHFR. The *TL* results from the HsDHFR seed could therefore be used as a baseline for identifying any benefits from using a parasite seed group. An improvement in the learning curve deficiency in comparison to the curve from the HsDHFR seed can therefore be interpreted as a positive effect directly relating to the seed parasite-protein strain. Any parasite seed benefits may relate to similarities in the protein structure of the seed and target, and this may especially be evident in detection of the rare category compounds.

In general, there were strong transfer effects from Pv and Pf seeds to assays containing other plasmodium strains in comparison to the Hs baseline; the only exception to this was with PfRdhfr (which again shows how difficult this target is).

- TcDHFR-active seeds showed a positive effect when used against assays containing Pv, Pf and PvRdhfr strains.
- The TcDHFR target benefits from having Pv or PfDHFR hits in the seeds.
- TbNMT-active seeds had a positive effect on identifying compounds active against PvDHFR and PvRdhfr, and also improved rare category selection for these targets.

These transfer simulations provide only a snapshot for each seed/target combination but they have begun to suggest patterns in seed/target strain similarity that could be exploited. Further experimental work would need to be undertaken to show the statistical significance of these effects.

Fuller analyses of the effects of active seeds applied to individual target strains are given in Appendix B.6.

5.6 Conclusions

The five AL strategies displayed a variety of features:

Active k-optimisation (described in section 4.5.2)

This algorithm has strong overall performance when selecting generally active compounds. The strategy of sampling unexamined areas of the chemical space allows rare category compounds to be found at a steady pace.

The main drawback of this strategy was its need to recompute ever larger matrix products at each cycle; growth measurements relative to the human orthologue are central to identifying the next best compounds to test, and needed to be continually re-evaluated. This feature was designed out of the algorithms subsequently developed that used simpler activity classification; however, it should not be considered a major practical problem as good performance will generally be more important than physical computational requirements (which can generally be solved by investment in more processing power).

SimplyGreedy (described in section 4.5.3)

Constructed as a baseline learning strategy using simple, greedy searching based on similarity to existing active classified compounds, this strategy gave good performance for general active compound selection. The process of only identifying active compounds which are similar to previously active compounds defines the algorithm as having minimal ability to identify rare compounds; this allows it to be used as a benchmark for rare category detection in other methods.

Preclustering (described in section 4.5.4)

These strategies demonstrated the power of classification of multiple negative/ inactive instances; in addition to using active results, they also use the likelihood that unknown compounds that are structurally similar to inactive seeds will also be inactive, thereby allowing them to be relegated to a later stage in the screening process. This promotes compounds with no significant similarity to those already examined, and potentially enables the earlier examination of rare active compounds.

The ability to find generally active compounds was slightly weaker when straight preclustering approaches were used, compared to the above strategies.

Transfer Learning (described in section 4.5.5)

This approach showed a significant improvement in detecting generally active compounds, owing to the higher prior likelihood of the initial seed group containing a large number of strong candidates. In some cases this will also relate to similarities between the parasite-protein structures of the target and from the transferred entity (e.g. PfDHFR and PvDHFR).

Rare category detection was improved when compared to *SimplyGreedy*.

Transfer Learning with Preclustering (described in section 4.5.6)

Combining the benefits of the *preclustering* and *transfer learning* algorithms gave very good results for the rate of selection of generally active compounds and for rare category detection; in keeping with the other classification-based methods, it also had the additional benefit of requiring a lower computing capacity than the prototype *active k-optimisation* method. If this method were to be developed further, it would be expected that modifications could be made to the final phase to improve detection rates for active compounds.

General comments

Further modifications could be made to the seed and unknown compound data sets to routinely relegate problematic compounds (i.e. toxic, autofluorescent) as knowledge is built up screen by screen.

The ideas upon which the *preclustering*/rare category detection are based could also be used in combination with either *active k-optimisation* or *SimplyGreedy*, especially if an artificial endpoint were desired. The operator would simply need to predetermine the point at which to run a routine based on proximity of prior inactive compounds to the remaining unknowns.

Chapter 6

Development of an econometric model of drug discovery

Adventures with yeast, part 6 of 7: The perfect Marmite sammich

2 slices	*Khorasan bread*
1 knob	*good quality butter*
1 measure	*Marmite*
1 bottle	*Cwrw Cawrfil*

Butter each slice of bread in a liberal fashion. Smear a sufficiency of Marmite onto one slice, and layer the sandwich.

Decant the Cwrw Cawrfil into a globed beer glass, taking care to leave the lees.

Sit back and watch the sea roll in.

An econometric model was developed to determine the utility of Eve for drug discovery; it identifies when conditions exist for efficiency gains, and when economic advantages might be found for active compound selection when compared to a linear screen. Assays that provide large numbers of potentially active compounds (e.g. PvDHFR) show the potential for efficiency gains when using AL selection strategies, whereas for targets that are difficult to hit (e.g. PfRdhfr, LmDHFR, SmDHFR), the econometric model shows that the AL strategies struggle to show any economic benefit.

The expected general advantages using the prototype *active k-optimisation* strategy and *SimplyGreedy* were readily shown; the additional advantages provided by transfer learning methods were also clearly seen in comparison to strategies based on internally-provided seed data.

The limitations of the model have been examined and discussed in terms of any assumptions made and limitations due to minimal information on the cost-benefit of the drug discovery phase in the public domain.

The potential beneficial effect of stronger rare compound detection, as offered by the *preclustering* strategies, could not be quantified. In order to accomplish this, a suitable range would need to be placed on the value of rare scaffolds compared to the active compounds that are similar to earlier seeds. Earlier detection of rare category compounds might allow screens to be terminated earlier, containing econometrically optimised information; if the value of rare compounds was known then experiment termination thresholds could be investigated for these strategies.

6.1 Background

There is little information in the public domain about the true cost of drug discovery and development. There are no published details that can be considered a gold standard (**Morgan *et al*., 2011**), with a large range covering estimates ($92m - $883.6m) based on 13 publications up to 2009. The available data is broken into discovery and development steps, and tends to show an approximate 1:2 split in costs across these tasks.

A breakdown of the discovery and development steps (**Paul *et al*., 2010**) provides the most recent cost estimates and attrition rates of candidate compounds, based on data provided by Eli Lilley & co. The target-to-hit and hit-to-lead discovery steps account for 9% of the discovery/development process.

Robot Eve's existing operational configuration is designed to fit into the target-to-hit region of the drug discovery process. The multi-target assays are expected to provide information on pathway-specific activity, with additional comparison against the equivalent human strain. The *active k-optimisation* strategy is designed to select suitable candidates and to explore based on these Hs-parasite growth differences. In the longer term, across several screens, these data could also be used to build up candidate-specific knowledge that highlights repeat offenders in terms of cytotoxicity and autofluorescence; this would assist cherry-pick screening by relegating these as unlikely candidates.

When considering libraries of existing drug therapies such as the JHCCL, the cost emphasis will be skewed by the additional knowledge available. It is envisaged that repositioning an existing therapy might be significantly less expensive (**Boguski *et al*., 2009**); arguably, a lead compound could progress far more rapidly to a Phase II trial (drug effectiveness) thanks to previous pre-clinical and Phase I (safety) studies once *in vivo* performance has been established. However, some structures might not respond well enough and only be considered lead compounds that require additional development, thereby reverting to the 'lead optimisation' step.

6.2 Econometric modelling for Eve

An econometric model for the differential advantage of using intelligent screening is displayed below; it was developed to determine the utility of Eve for drug discovery, i.e. the range of conditions for which using a Robot Scientist to guide candidate compound selection is economically advantageous compared with performing a standard whole-library screen. In all types of screen there are costs associated with physically using up the contents of the compound library, as well as utility, time and labour costs; in mass screens the cost of loss of compound is significantly lower than in more complex confirmation screens.

$$\Delta\, Utility\ of\ Eve = \sum_1^{Nm}(Tm + Cm) + \sum_1^{Nx}(Tc + Cc - Uh) + \sum_1^{Ne}(Tm - Tc + Cm - Cc) \quad (12)$$

Nm	-	Number of compounds not assayed by Eve
Tm	-	Cost of the time to screen a compound using the mass screening assay, $
Cm	-	Cost of the loss of a compound in the mass screening assay, $
Nx	-	Number of hits missed by Eve
Tc	-	Cost of the time to screen a compound using a cherry-picking (confirmation or intelligent) assay, $
Cc	-	Cost of the loss of a compound in a cherry-picking assay, $
Uh	-	Utility of a hit, $
Ne	-	Number of compounds assayed by Eve

The net utility is made up of three cost components; the cost of:

(i) not screening Nm compounds which, based on the QSAR learning, are less likely to be hits [i.e. a cost saving];

(ii) not finding any hits (Nx) that might be present in this unscreened set ($Uh >> Tc + Cc$) [an indeterminate, negative, opportunity cost];

(iii) cherry-picking Ne compounds, ($Tc + Cc$)>($Tm + Cm$) [physical negative cost].

6.3 Application of the model to Active Learning curve simulations

6.3.1 Model rearrangement

Rearranging the utility equation assists with simplifying the calculations when using it to evaluate the AL simulation loops:

$$\Delta\, Utility\ of\ Eve = \sum_1^{Nm+Ne}(Tm + Cm) - \sum_1^{Ne-Nx}(Tc + Cc) - (Nx \times Uh) \qquad (13)$$

The first term is now effectively the full cost of performing a standard library screen. For Eve and the Maybridge HitFinder library, upfront and ongoing library and consumables costs for a library screen are fixed.

The second term is the operational cost associated with cherry-picking Ne compounds but not the Nx undiscovered hits when the Active Learning algorithm is selecting candidate compounds based upon the QSAR learning.

The third term is the economic value of a hit compound against the parasite (Uh), and the number of such hits (Nx) yet to be found by the Active Learning; again, this opportunity cost is recalculated in the simulations for each loop in the screen.

6.3.2 Econometric modelling using simulation data

The *active k-optimisation* AL algorithm was applied to the seed input data and the unknown compound SMILES codes; simulated learning curves were produced for each parasite strain using the proxy confirmation data (see section 5.1). The progression of these learning curves was then compared to the base case of a linear progression throughout the screen in accordance to the utility equation (Equation 12). For each 96 compound loop, the number of proxy confirmed hits and compounds screened to date (Ne) were applied to the utility equation, together with the fixed utility and cost terms. An example of the resultant 2D plot for the PvDHFR strain is shown in Figure 6.1.

	Compounds	Screen hits
Full screen	14386	316
Seed	958	25
Unknowns	13428	291

No. of loops	No. of screen hits in simulation loops						
	n=0	n=10	n=20	n=30	n=40	n=50	n=60
n+1	8	53	106	133	162	190	204
n+2	12	60	109	137	163	192	206
n+3	17	67	111	141	167	194	208
n+4	23	73	112	142	169	196	210
n+5	29	79	116	144	174	198	214
n+6	30	86	117	146	179	199	216
n+7	36	90	118	148	182	200	217
n+8	39	95	123	152	184	201	218
n+9	44	98	125	154	186	202	221
n+10	47	104	127	157	189	203	223

No. of loops	No. of screen hits in simulation loops						
	n=70	n=80	n=90	n=100	n=110	n=120	n=130
n+1	226	244	260	279	291		
n+2	228	247	263	280			
n+3	229	250	264	281			
n+4	231	251	268	283			
n+5	233	252	271	284			
n+6	235	253	273	285			
n+7	238	254	274	286			
n+8	240	255	275	288			
n+9	242	256	277	289			
n+10	243	259	278	290			

Table 6.1: Simulation data for the TS3 PvDHFR target - number of hit compounds found with progression of the loop count.

Note: the utility increase example displayed below is the percentage increase over the equivalent stage in the base case linear screen; this is true for all the utility curves shown in this document. Overall utility will become negative once the screen has progressed beyond the point at which the benefits from AL are outweighed by its additional costs.

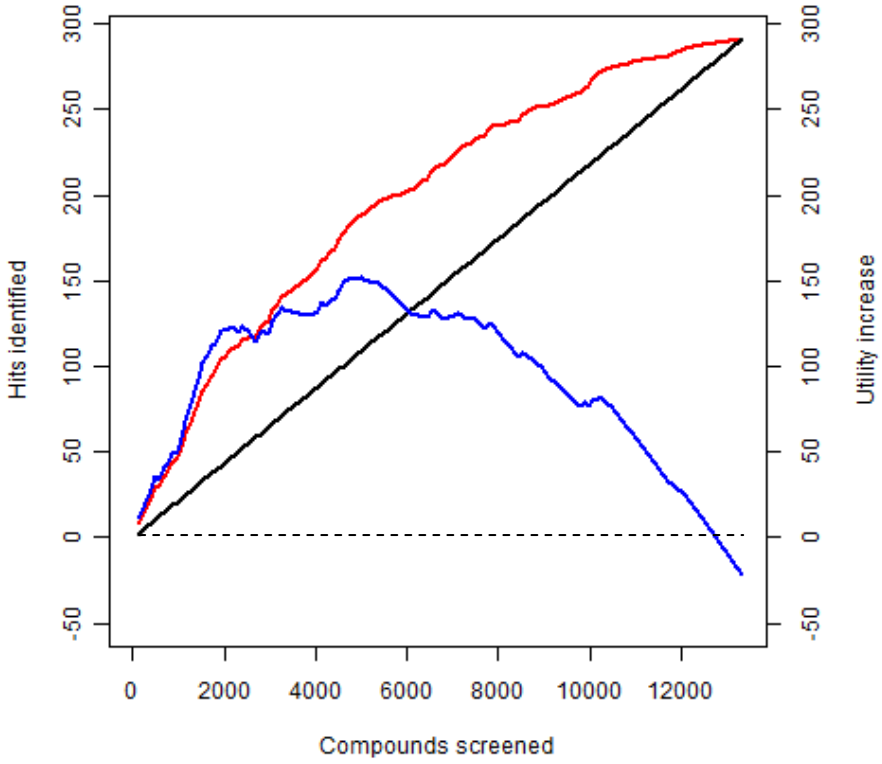


Figure 6.1: Hits found in TS3 PvDHFR simulation (red) versus base case (black), and the resultant econometric utility % increase (blue), based on data in Table 6.1

(Tm = 800 per cycle, Cm = 0.4 per compound, Tc = 1000 per cycle, Cc = 3 per compound, Uh = 2000)

The cost of a cycle of AL includes both the time and cost of the computing power, and the cost of testing a 96-compound batch in the cherry-picking assay. Based on the *active k-optimisation* simulated cherry-picking cycles, these terms were also calculated for examples of each parasite target (Figure 6.2).

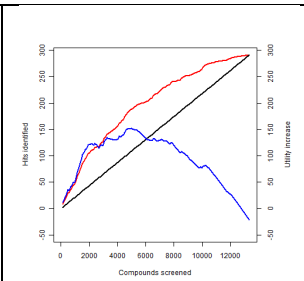
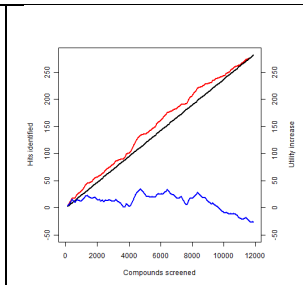
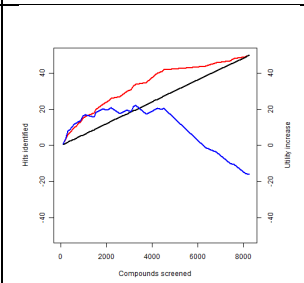
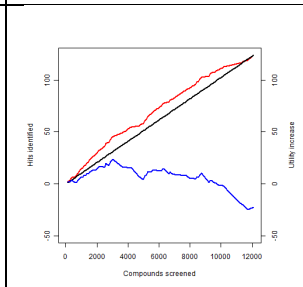
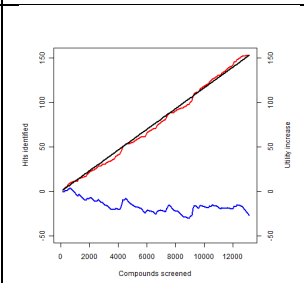
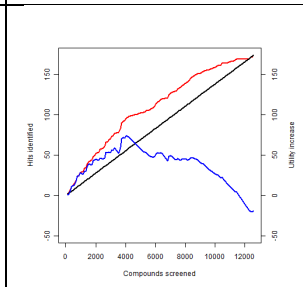
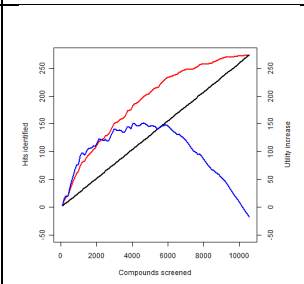
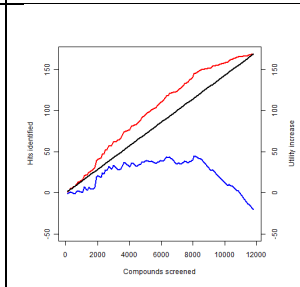
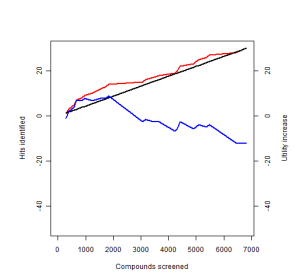
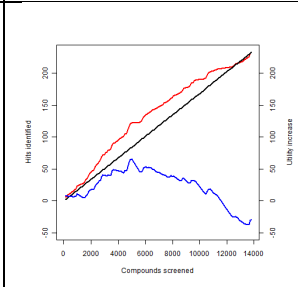
Screen	DHFR target		DHFR target	
TS3	Pv		PfR	
TS4	Tb		Sm	
TS5	PfR		Tc	
TS6	Pv		Pf	
TS7	Lm		PvR	

Figure 6.2: Simulations of intelligent screening for each DHFR target using the *active k-optimisation* strategy.

A utility landscape was also constructed for the TS3 PvDHFR parasite data across a range of ML efficiencies and drug economic values (Figure 6.3); the value of a hit compound for this study ranged from £2K to £15K, based on a broad estimate of the number of hits required to give sufficient drug-like lead compounds to commence lead optimisation studies. Variation in the time-cost ratio comparing mass to intelligent screening (Tc/Tm) was studied, and utility versus cost of compound loss during cherry-picking (Uh/Cc) was also evaluated.

Efficiency benefits are more readily found when the time-cost ratio tends towards unity, and when the utility of a hit increases compared to the cost of loss of library compounds. These observations would be expected of an AL strategy capable of improving the selection process for drug-like candidates.

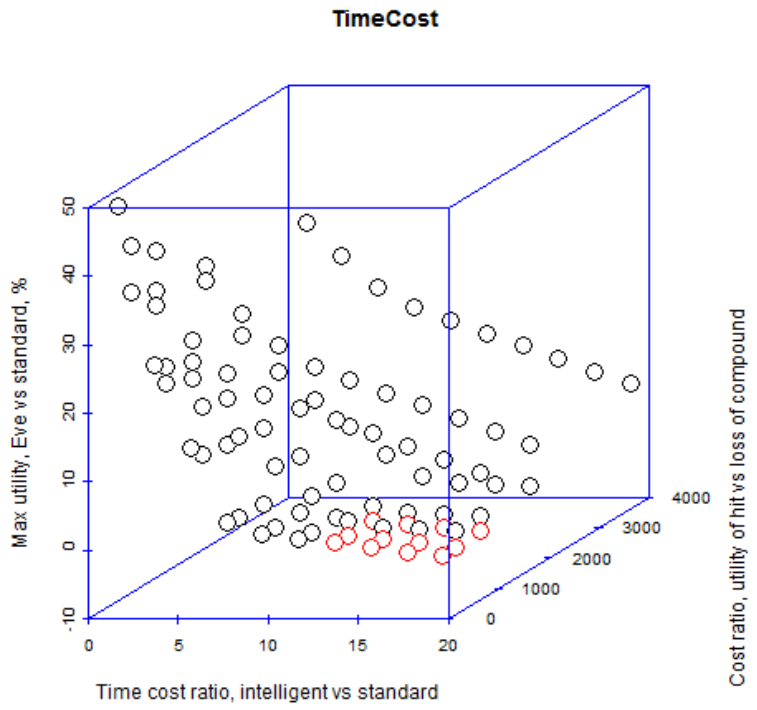


Figure 6.3: Utility landscape for TS3 PvDHFR, Time-ratio = Tc/Tm, and Cost-ratio = Uh/Cc

Figure 6.4 displays the effect of the econometric model when applied to example SimplyGreedy simulations. Figure 6.5 depicts a similar set of examples from the Transfer Learning/ Preclustering (TS > 0.40) simulations based on the compounds previously identified as active versus PfDHFR in TS6.

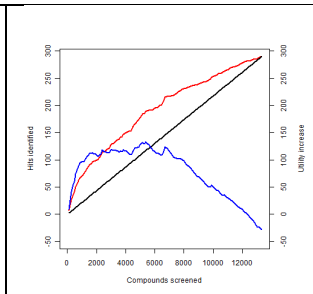
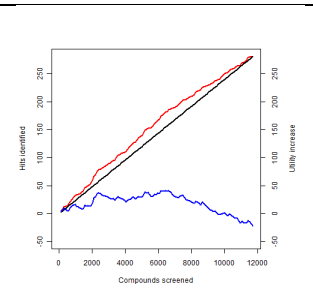
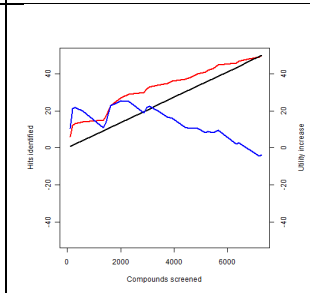
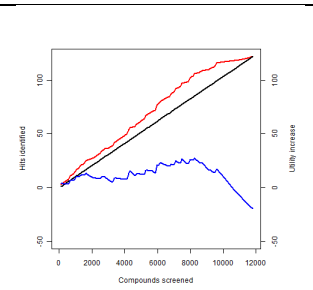
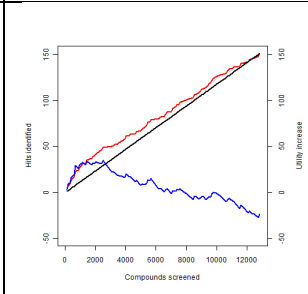
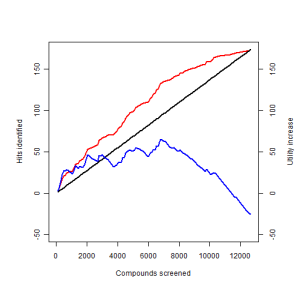
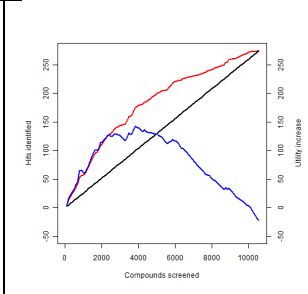
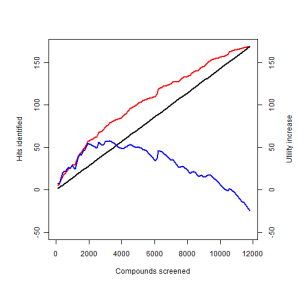
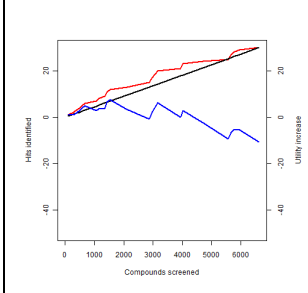
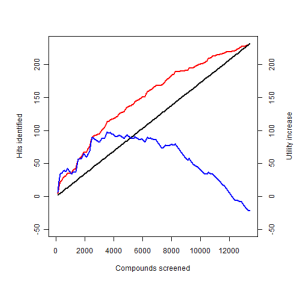
Screen	DHFR target		DHFR target	
TS3	Pv		PfR	
TS4	Tb		Sm	
TS5	PfR		Tc	
TS6	Pv		Pf	
TS7	Lm		PvR	

Figure 6.4: Intelligent screening simulations, DHFR targets, SimplyGreedy

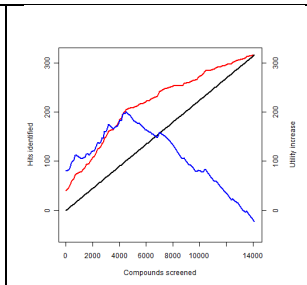
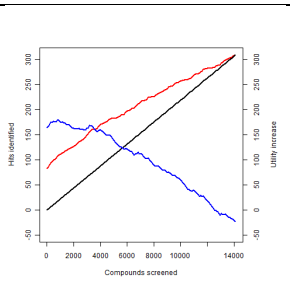
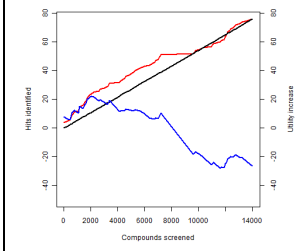
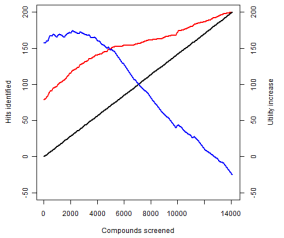
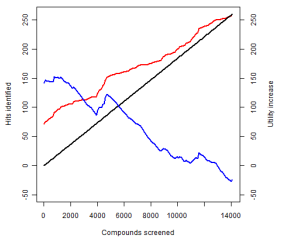
Screen	DHFR target		DHFR target	
TS3	Pv		PfR	
TS4	Tb		Sm	
TS5	PfR		Tc	
TS6	Pv		Pf	
TS7	Lm		PvR	

Figure 6.5: Intelligent screening simulations, DHFR targets, Transfer Learning from PfDHFR with preclustering at TS > 0.40.

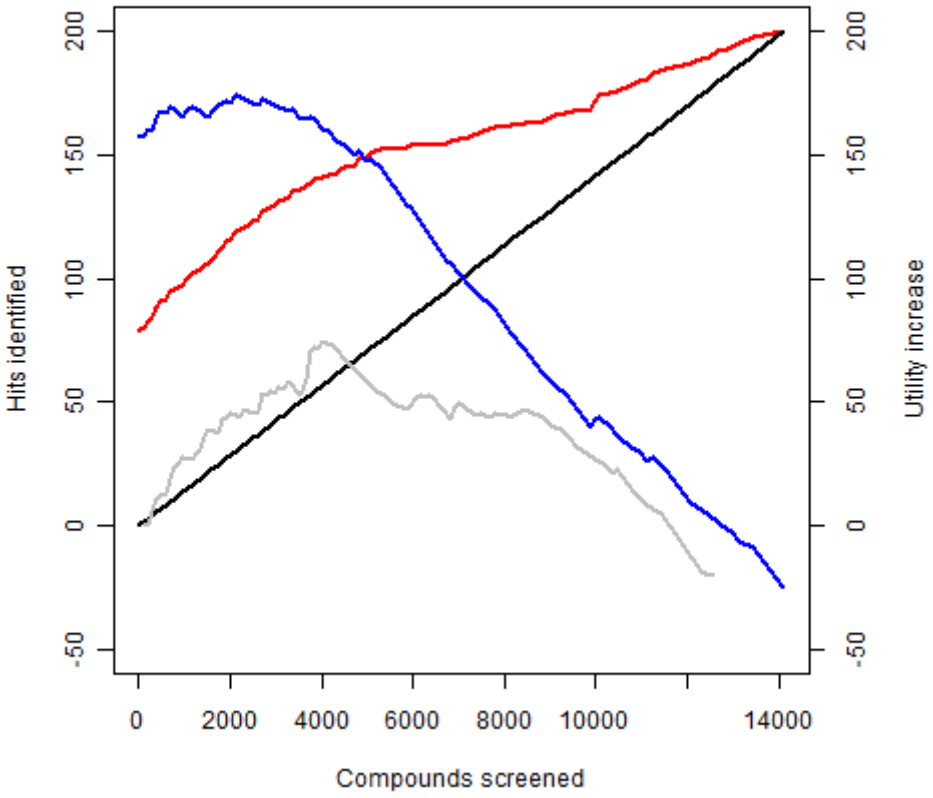


Figure 6.6: Intelligent screening simulations, TcDHFR targets: Hits identified by Transfer Learning from PfDHFR with preclustering at TS > 0.40 (red) versus random selection (black). Relative econometric performance (utility increase) of *active k-optimisation* (grey) and transfer learning/preclustering (blue)

Figure 6.6 overlays the econometric curves for TcDHFR from Figures 6.2 (grey) and 6.5 (blue), showing the strong performance offered by the combination of Transfer Learning and Preclustering (TLP) by its effect on overall utility. The initial boost to utility enables the curve to remain above the *active k-optimisation* curve throughout the experiment. It should be noted that the TLP approach will also have provided stronger performance in rare category detection.

6.4 Modifications to the econometric model

One aspect that has not been considered in the econometric model is whether the utility of a hit (Uh) has a variable value. It has been argued in section 5.5 that rare category active compounds are likely to be of significant interest, as their structures are dissimilar to other active compounds. The ability to promote and find these candidate compounds is considered a potentially important feature of an AL algorithm for Robot Eve. The ability of the *active k-optimisation* strategy to search for diverse compounds meant that it performed reasonably well by finding rare category compounds linearly throughout the simulation. Introduction of the *transfer learning with pre-clustering* strategy gave a strong improvement on detection of these rare category compounds.

6.5 Discussion

6.5.1 The econometric model helps to describe the efficiency gains when using Active Learning. Although the applied financial variables are not directly taken from the pharmaceutical development process, they are considered to be fair guesses and Figure 6.3 displays the effect of a spread of these values. In Figure 6.3, utility is improved as the time cost of intelligent screening approaches that of standard mass screening, and is also improved with increasing value/utility of hit versus the cost of loss of compound (raw material replenishment).

6.5.2 As expected, there is little difference in the econometric performance of the prototype *active k-optimisation* strategy and *SimplyGreedy*. The curves for those targets with a strong set of active compounds are broadly similar using either strategy, as indicated by the earlier deficiency measurements.

The poor quality of the curves for PfR, Sm & LmDHFR suggest that these parasite proteins are difficult targets to fit into an econometric model. This reinforces the opinion that these are difficult drugable targets, as identified by the low number of active compounds found during the mass and confirmation screens (Chapter 3 and Appendix A.6).

6.5.3 The positive effect of transfer learning on econometric performance is clear. The econometric model has only been applied to a few of the transfer learning curves, but the expected advantages are easily highlighted. Figure 6.6 shows an example of the large boost to econometric performance when comparing the overlaid transfer learning curve (blue) and the prototype curve (grey) for TcDHFR. Similar strong effects would also be seen with the other targets.

6.5.4 There remains the question of when to terminate the screen. The simulations and econometric model suggest that there are advantages in running a partial screen if there are capacity limitations on the equipment. The ability of any Active Learning regime to highlight active rare category compounds would provide additional impetus to curtail the screen after their identification, if these are of high value as expected.

Rare category compounds were defined by having no near active or inactive neighbours in the Maybridge Hitfinder library; this might be impractical in a large commercial library, so it is suggested that a threshold quantity of near inactive neighbours could be established. It is also suggested that the current tail set of compounds with near inactive neighbours might provide more valuable information if the set were graded for examination based on a rising proportion of near inactive neighbours.

6.5.5 The model has certain limitations, as there is little information in the public domain that describes some of the values applied to its parameters. Assumptions have been made for the utility of active candidates in the drug discovery phase, and only costs associated with the experiments in Aberystwyth could be used. The model can readily be used for comparing economic merits of different simple AL approaches reported herein, but might struggle if used to provide an absolute description for HTE in the wider pharmaceutical industry.

Chapter 7

Conclusions and further work

Adventures with yeast, part 7 of 7: Gwynant Cochnant

4.0 kg	*Maris Otter pale malt*
0.2 kg	*crystal malt*
0.1 kg	*dark crystal malt*
0.05 kg	*light chocolate malt*
50 grams	*Challenger hops*
50 grams	*Mount Hood hops*
1 packet	*SafaleS-04 beer yeast*
1	*Irish moss tablet*

Add the grains to the mash tun with 20 litres of water. Steep at 70-75°C for two hours. Remove the grain sack and allow it to drain into the wort; sparge it with boiling water until the sugars have been depleted. Remove 500 ml of wort and use it to make a yeast starter. Bring the bulk wort to the boil; maintain the volume at approximately 25 litres. Add the Challenger hops when the wort is at a rolling boil, and boil for 60 minutes; add 25 grams of Mount Hood hops at 30 minutes from the end of the boil, and a further 25 grams at 15 minutes from the end, together with the Irish moss tablet.

Cool the finished wort; decant it to a fermentation vessel and add the yeast starter. Record the original gravity (~1040), and the final gravity (~1010 after 10 days). Prime the finished brew with 100 grams of white sugar in 1 litre of boiled/cooled water prior to bottling or casking.

This recipe is a "work in progress" for the purpose of revitalising Bragdy Gwynant, the world's smallest commercial brewery (according to Guinness World Records).

The scope of this thesis was very broad, encompassing the need to understand and develop several computational methods that explored aspects of drug discovery and design, biochemical pathways, and general data analysis. Whilst much of the early focus was on the empirical aspects of the Robot Scientist programme, analysis of Eve's output was a necessary hurdle to overcome before embarking on the development of Active Learning processes. The resultant data sets have allowed an independent analysis of the similarities between active, drug-like compounds.

The simple input/output data structures required by the AL methods developed herein would readily lend themselves to other drug discovery processes. Similarly, the use of inactive labels, as shown in the *preclustering* strategy, could be adopted for the exploration of other problems where the detection of rare categories is important e.g. detection of financial fraud or network security breaches.

7.1 Primary achievements

7.1.1 A robust data analysis process has been built for Robot Scientist Eve

The method developed to analyse mass screen data for Eve incorporates several descriptors from across the full growth curve of the substrate, and has shown itself to be robust after application across many mass screens. It has shown evidence of better sensitivity versus an alternative method based on measurements of final growth alone, which was retrospectively compiled after completion of several screens. Now that a larger body of experience and confirmation data has been acquired, it would be feasible to tune the mass screen analysis to reduce noise caused by repeat instances of autofluorescent and cytotoxic compounds, although these are very few in number relative to the bodies of the libraries in use. Similarly, a better background signal might now be possible to build, by using a moving average for individual compounds in additional to the in-plate negative controls.

Similarly, the processes for analysing confirmation data worked well, with the decision tree rules providing a satisfactory indicator of borderline activity. It might be useful to enhance these rules with an additional indicator to give the magnitude of activity, although this is already provided in a semi-quantitative fashion using a scoring process.

7.1.2 The ability of Eve to consistently find active compounds, and perform quantitative studies of selected compounds, was demonstrated

Eve's library screening studies have identified a large number of molecules in the Maybridge Hitfinder collection that have confirmed activity versus the parasite strains. These instances are recorded in Appendix A.8.

There are also several compounds with confirmed activity originating from the Johns Hopkins Clinical Compound Library; these instances are recorded in Appendix A.9.

7.1.3 A prototype Active Learning system was implemented

The prototype *active k-optimisation* algorithm was applied to Eve's mass screen data for end-of-test growth to provide an Active Learning regime. When combined with the above data analysis process, Eve was able to operate a closed loop drug discovery process.

Due to cost and time constraints only three rounds of AL were applied; the output from these was used to identify a means whereby mass screen data might act as a proxy for confirmation screen data, thereby allowing an AL cherry-picking simulator to be constructed.

7.1.4 Alternative Active Learning processes were developed

The *active k-optimisation* strategy was benchmarked against a greedy learning algorithm based on Tanimoto Similarity between candidates and active compounds; this *SimplyGreedy* algorithm was based on classification rules rather than quantified activity, and was further developed using transfer learning techniques to give a simple yet effective mechanism to identify active candidates.

7.1.5 Strategies for detecting rare category compounds were developed

The *SimplyGreedy* algorithm and its off-shoots were combined with clustering techniques to explore the compound library. This led to the identification of similarity thresholds below which active compounds are unlikely to be found efficiently, and enabled algorithms to be built which switched to searching unexplored spaces in the library. The resultant rare category detection algorithms allowed Eve to promote active compounds that would otherwise have been difficult to find until later stages.

By allowing an experiment to terminate earlier, these compounds are inevitably now found at the expense of less 'rare' compounds, but compensate by providing diverse chemical structure scaffolds.

7.1.6 Generalised applicability of Eve's Active Learning strategies

The Active Learning strategies based on classification of drug-like activity have been developed successfully. The simple input data structure means that these strategies should be readily adaptable to other drug discovery regimes, and will also have applicability for dealing with other problems where rare category detection is required.

The inputs mainly used in this thesis (i.e. FPT2 fingerprints, Tanimoto Similarity, and activity classification data) are fairly simple information sets but, for drug discovery work, these Active Learning strategies could easily adopt other methods of showing potential biochemical similarity between candidate compounds. The main aspect that would need to be considered for using other information sources would be the similarity thresholds at which the activity measurement system can give a rich vein of candidate compounds.

7.2 Secondary Achievements

As might be expected of any major research and development programme, several discoveries have been made that were outside the original scope of the project:

7.2.1 The benefits of discrete, classification-based inputs over the continuous variables for end-of-test growth quantification were shown

The *active k-optimisation* strategy provided a strong selection process, with good overall identification of active compounds together with an effective means of searching for rare category compounds throughout experiments. However, this algorithm required much more computational capacity than the simpler, classification-based algorithms developed from the initial *SimplyGreedy* strawman, which in turn were eventually constructed to provide both strong overall performance and very strong rare category detection. The latter methods will have benefitted from using transfer learning, and it would have been of interest to apply a similar

approach as part of the seed for *active k-optimisation* had there been appropriate facilities for such work.

7.2.2 Benefits were shown when classifying activity using several attributes of the yeast growth curve, in comparison with end-of-test quantification rules

When the initial data analysis rules (set up using multiple attributes of the yeast growth curve) were compared with simple growth-based rules, they showed strong performance in identifying a wider range of active candidates. These rules performed better owing to their ability to identify candidates that might otherwise have been miss-labelled as false negatives; simple end-of-test growth rules were found to misread candidates with slower growth rates or inhibited lag phases.

7.2.3 A possible mode of parasite growth inhibition was shown for some JHCCL candidates

Two of the JHCCL candidate compounds with confirmed screen activity have previously been suggested as having anti-malarial activity, and specific protein targets have been suggested in the literature. Follow-on *in vivo* evaluation of these compounds has suggested potential alternative sites for their anti-parasitic behaviour.

It is possible that Triclosan (SM-JH-10450) has been identified as an anti-folate for *Plasmodium sp*.; earlier indications of its anti-parasitic behaviour (**Surolia and Surolia, 2001**) suggested its activity was through inhibition of FabI, its bacterial target (**McMurry *et al*., 1998**), although it was later convincingly shown that FPfFabI is not the main target of the anti-*Plasmodium* activity of triclosan (**Yu *et al*., 2008**).

The anti-cancer angiogenesis inhibitor Tnp-470 has previously been reported to possess activity *in vitro* against *P. falciparum* - with the target believed to be methionine aminopeptidase 2 (**Arico-Muendel *et al*., 2009**). Eve found weak activity against *P. falciparum* DHFR but much stronger activity against *P. vivax* DHFR; this suggests that, similar to triclosan, it may be attacking the DHFR target.

7.2.4 A family of Maybridge compounds having activity versus the PvDHFR target was investigated

Twelve candidates (5 from the Maybridge Hitfinder library, 7 from the full Maybridge collection) that had strong similarity to a small group of active seed compounds were identified for confirmation testing (see section 3.3.4). Seven were found to be active versus the PvDHFR target, with two inactive and three untested. The inactive compounds were smaller molecules with no large functional groups attached, and could be described as small molecules for drug lead development in this context.

This family might be a useful starting point for a lead compound versus PvDHFR.

7.2.5 The economic benefits of Active Learning were shown

An economic model was developed to show the economic benefits of AL in the drug discovery process. Comparison between the prototype *active k-optimisation* strategy and a simple, linear candidate selection process showed the advantage of informed selection, and that significant efficiency gains are possible.

The same set of conditions was used to test a simple, greedy AL algorithm, together with a number of alternative selection mechanisms for candidate compounds. The overall shape and measured deficiencies for the other non-transfer AL algorithms suggest that similar econometric performance would be found. An exception to these observations might be found for the transfer learning algorithms, where the learning curve is strongly boosted by many early active examples.

It has been mooted that the intrinsic value offered by finding rare category compounds would boost the overall econometric value of a screen; the ability to find such items at an earlier stage might allow the search to be truncated, with the screen value deemed as maximised. The combined transfer learning/preclustering strategy was very successful in promoting compounds that were otherwise difficult to find; in most cases >90% of the rare category compounds were detected by the end of the second phase of this strategy, and would therefore have much higher partial deficiency values at this point compared to other strategies.

7.3 Further Work

Eve's ML rules could be extended to include some of the ideas used in Cambridge's proposed data analysis systems, e.g. finding and removing superactive (toxic and autofluorescent) compounds, removing consistently problem wells, relegating promiscuous active compounds.

If experience across more screens were available it might be possible to build rules to test whether the selection strategy is working effectively. Empirical evidence suggests that some targets are difficult to hit, and there might need to be a larger body of active compounds needed to assist with predictions. In such cases, a switch to an alternative selection strategy might be beneficial, or even a switch back to mass screening mode to acquire a broader data set.

It is suggested that SMILES FP2 fingerprints based on fragments of 7 atoms (Daylight 0/7 configuration) are less productive in QSAR studies than larger fragments (up to 10 atoms in length, Daylight 3/10 configuration) (**McGaughey *et al*., 2007**). The rcdk toolkit allows for changes in chain length, and such effects could be investigated prior to completing a publication based on the AL studies.

Finding the dividing line between potentially active compounds and toxic ones is an interesting subject for Eve, and one that probably requires different approaches for the Maybridge Hitfinder and JHCCL libraries. There were occasions where JHCCL compounds were considered borderline toxic when using the methods set up with Maybridge data, but could well be useful for further evaluation.

If a lead originates from the approved drugs in the JHCCL, how much more valuable is it than a compound from a standard library such as the Maybridge collection? Could different, less strict toxicity criteria be constructed for approved drugs in Eve's decision tree rules, compared to those with less readily-available information?

An enhancement of the Robot Eve programme would be to expand the libraries to include a larger set of existing drug therapies. Similarly, other libraries outside of normal pharmaceutical collections (e.g. the SIDR Natural Products Collection (**Harvey *et al*., 2010**)) might lend themselves to exploitation by Eve's processes and targets, in accordance with the ideas for niche research areas for academia (**Lipinski, 2006**).

Appendix A

Experiment results for Robot Scientist Eve

A.1 Assay strains

Screen	Target	Strain/Fluorophore		
		mcherry	sapphire	venus
TS3	DHFR	Hs	Pv	PfR
TS4	DHFR	Hs	Tb	Sm
TS5	DHFR	PfR	Hs	Tc
TS6	DHFR	Hs	Pv	Pf
TS7	DHFR	Hs	Lm	PvR
TS8	DHFR	Hs	Sa	Sa
TS9	DHFR	Hs	Sa	Sa
PGK1	PGK	Hs	Sm	Tc
PGK2	PGK	Hs	Tb	Pv
NMT1	NMT	Hs	Tb	Pv
NMT2	NMT	Hs	Sm	Tc

A.2 List of mass screens

Assay	File ID	Compounds	
		Maybridge	JHCCL
TS3	MS_63_1_15_20110414203535.csv	14386+17	1249
TS4	MS_64_1_16_20110414205040.csv	14386+17	1245
TS5	MS_71_1_17_20110414210718.csv	14380+17	1249
TS6	MS_77_1_22_20110316174043.csv	14099	-
	MS_TS6_JHCCL.csv	-	1248
TS7	MS_80_1_25_20110630113217.csv	14376+17	1252
TS8	MS_83_1_26_20111028130222.csv	14224+17	284
TS9	MS_85_1_28_20120116103737.csv	13925+17	297
PGK1	MS_72_1_18_20110414212243.csv	14347+17	1249
PGK2	MS_74_1_20_20110414213751.csv	14272+17	1249
NMT1	MS_78_1_23_20110414215320.csv	14376+17	1250
NMT2	MS_79_1_24_20110414180141.csv	14376+17	1253

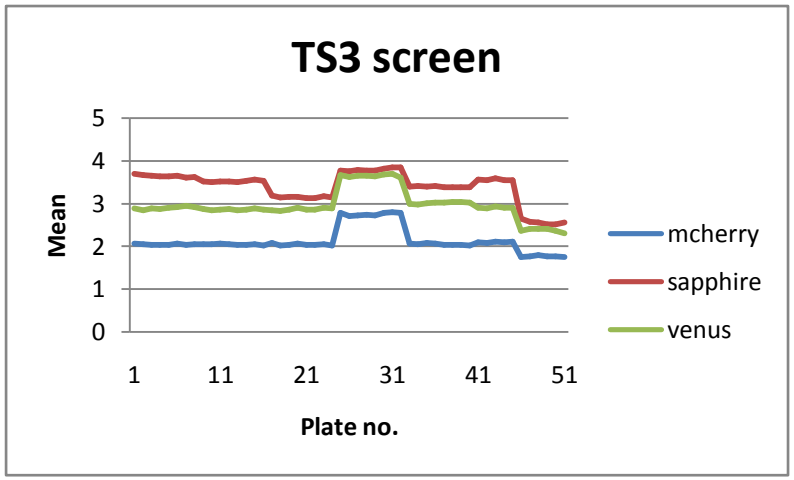
A.3 Mass screens, Maybridge Hitfinder activity

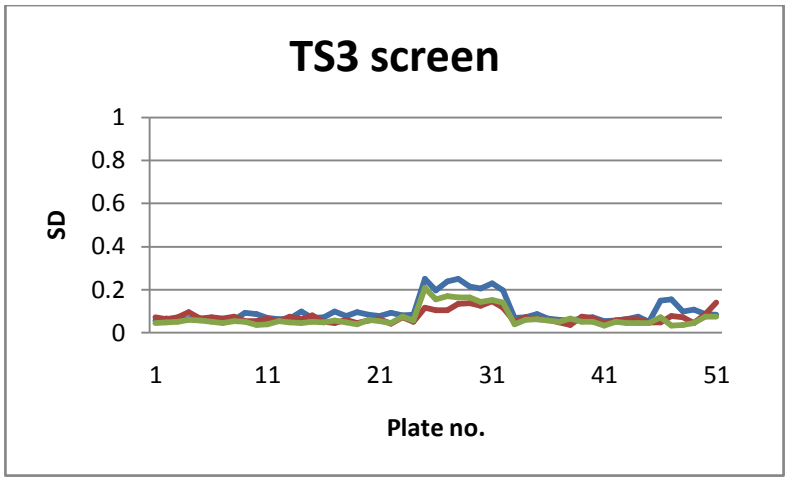
Assay	Target	Strain	Compounds	Mass screen active compounds			
				Clean	Co-hits	Possibly Toxic	Total
TS3	DHFR	Hs	14386	173	198	13	384
		Pv		236	59	21	316
		PfR		76	225	8	309
TS4	DHFR	Hs	14386	193	78	8	279
		Tb		34	30	11	75
		Sm		62	82	5	149
TS5	DHFR	PfR	14380	130	41	8	179
		Hs		6	19	12	37
		Tc		107	73	20	200
TS6	DHFR	Hs	14099	46	62	11	119
		Pv		203	81	19	303
		Pf		64	113	10	187
TS7	DHFR	Hs	14376	195	51	20	266
		Lm		8	31	16	55
		PvR		213	39	8	260
TS8	DHFR	Hs	14224				
		Sa					
		Sa					
TS9	DHFR	Hs	13925				
		Sa					
		Sa					
PGK1	PGK	Hs	14347	171	174	4	349
		Sm		12	16	9	37
		Tc		101	176	7	284
PGK2	PGK	Hs	14272	57	38	9	104
		Tb		13	17	7	37
		Pv		60	39	8	107
NMT1	NMT	Hs	14376	18	24	14	56
		Tb		200	98	13	311
		Pv		139	97	13	249
NMT2	NMT	Hs	14376	173	74	7	254
		Sm		29	36	5	10
		Tc		120	88	10	218

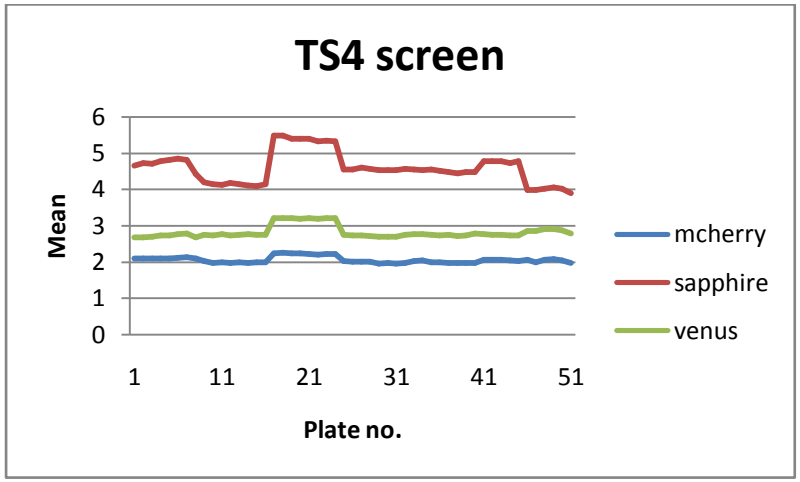
A.4 Mass screens, JHCCL activity

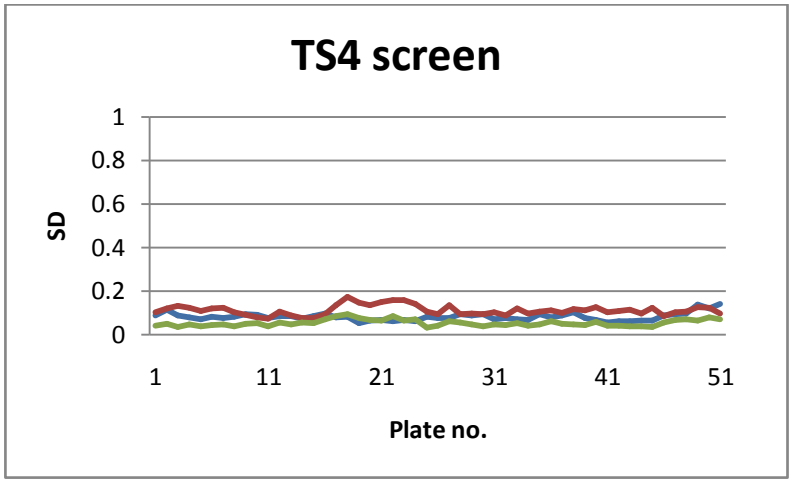
Assay	Target	Strain	Compounds	Mass screen active compounds			
				Clean	Co-hits	Possibly Toxic	Total
TS3	DHFR	Hs	1249	-	1	1	2
		Pv		7	3	1	8
		PfR		3	4	-	7
TS4	DHFR	Hs	1245	5	-	-	5
		Tb		2	1	1	4
		Sm		3	1	1	5
TS5	DHFR	PfR	1249	1	-	1	2
		Hs		-	2	3	5
		Tc		3	2	1	6
TS6	DHFR	Hs	1248	49	12	3	64
		Pv		7	7	1	15
		Pf		2	11	-	13
TS7	DHFR	Hs	1252	2	6	2	10
		Lm		2	7	1	10
		PvR		10	3	-	13
TS8	DHFR	Hs	284				
		Sa					
		Sa					
TS9	DHFR	Hs	297				
		Sa					
		Sa					
PGK1	PGK	Hs	1249	60	10	5	75
		Sm		1	1	4	6
		Tc		2	9	2	13
PGK2	PGK	Hs	1249	118	7	5	130
		Tb		2	2	3	7
		Pv		1	5	1	7
NMT1	NMT	Hs	1250	4	3	2	9
		Tb		4	8	-	12
		Pv		7	9	1	17
NMT2	NMT	Hs	1253	71	29	1	101
		Sm		4	6	1	11
		Tc		7	31	1	39

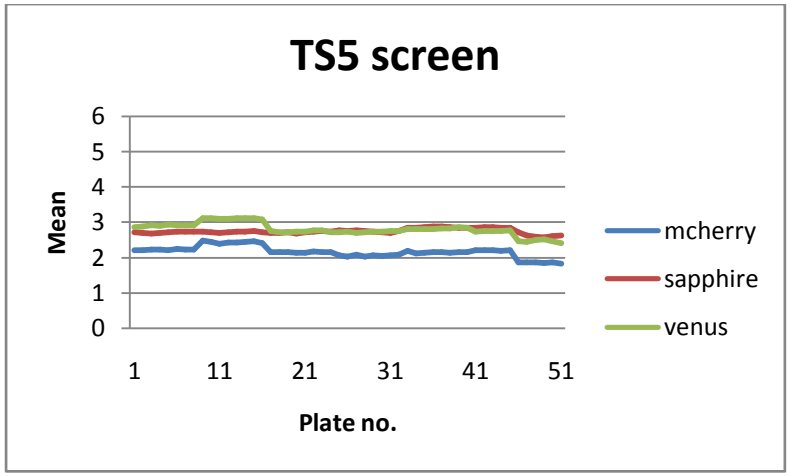
A.5 Mass screens: negative control statistics for doubling time

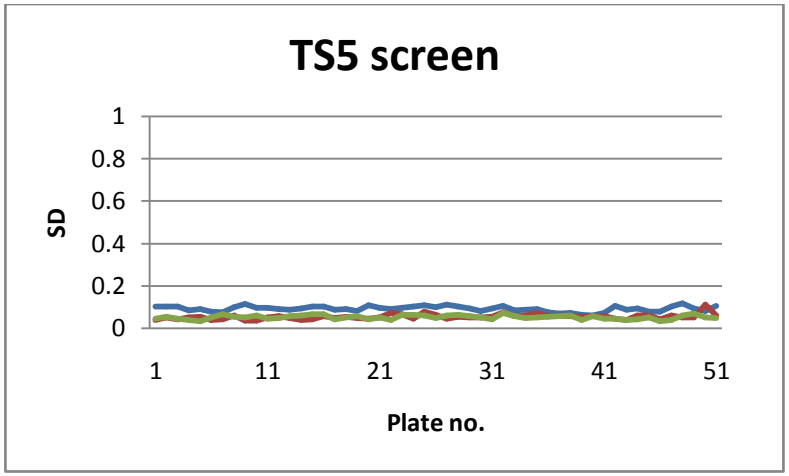


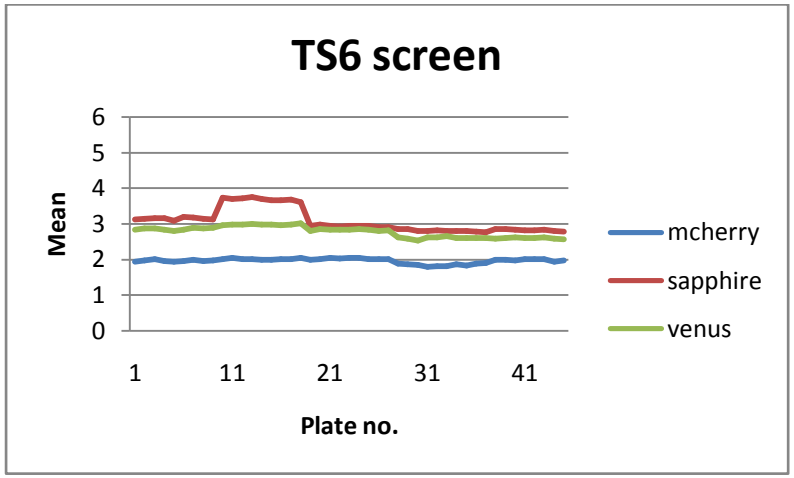


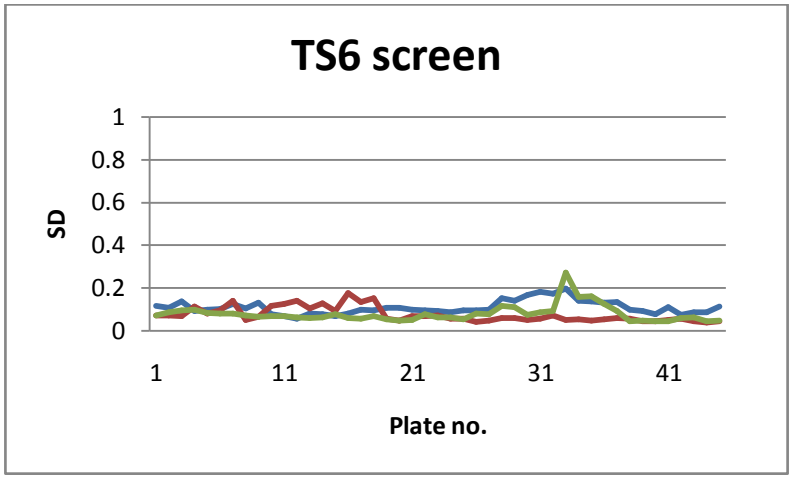


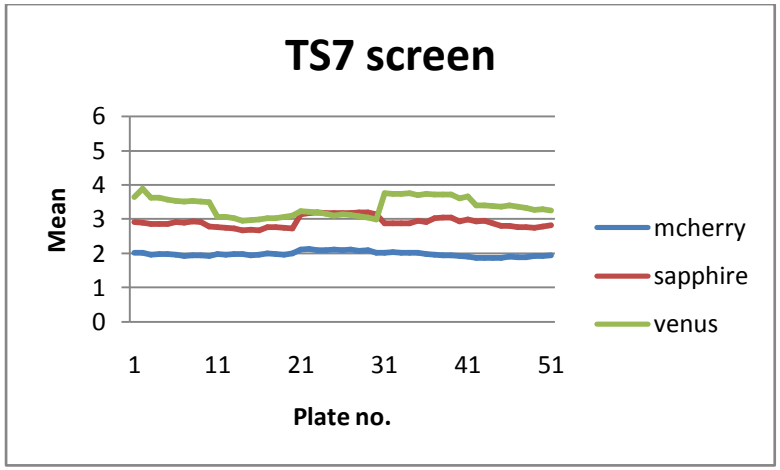


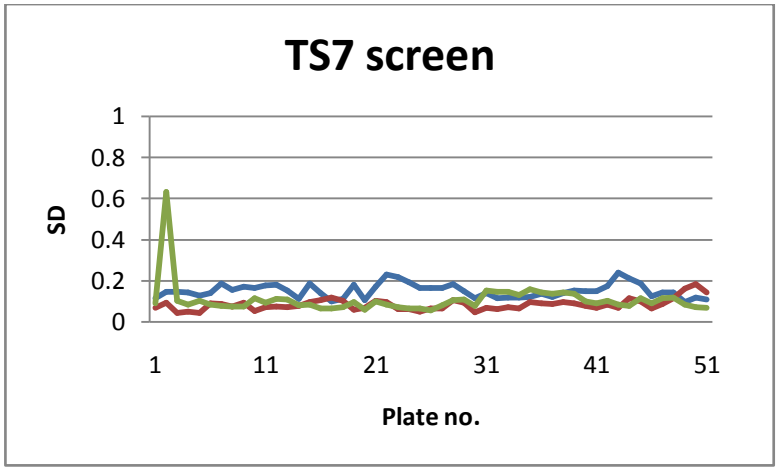

TS5 screen
Mean
6
5
4
3
2
1
0
1
11
21
31
41
51
Plate no.
mcherry
sapphire
venus

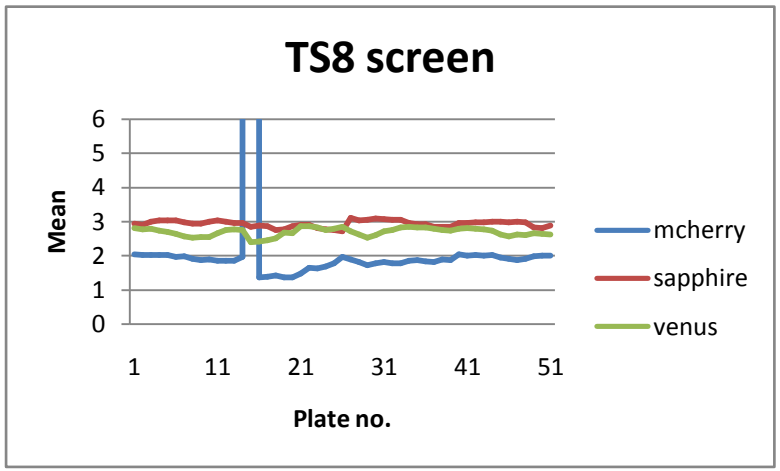

TS5 screen
SD
1
0.8
0.6
0.4
0.2
0
1
11
21
31
41
51
Plate no.

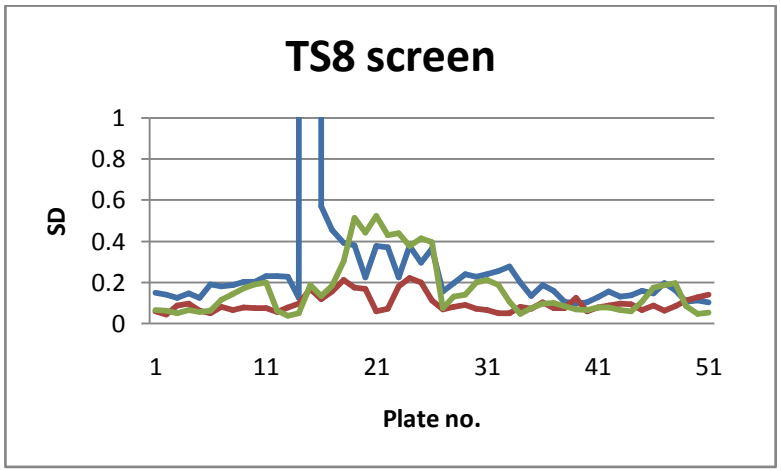

TS6 screen
Mean
6
5
4
3
2
1
0
1
11
21
31
41
Plate no.
mcherry
sapphire
venus


TS6 screen
SD
1
0.8
0.6
0.4
0.2
0
1
11
21
31
41
Plate no.

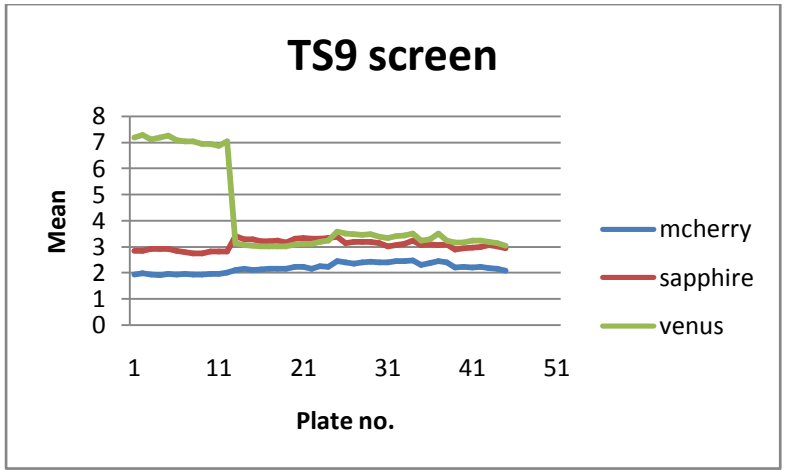


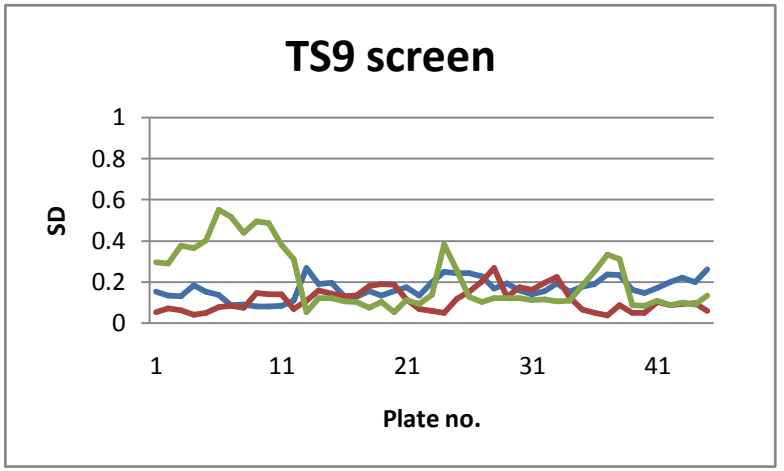


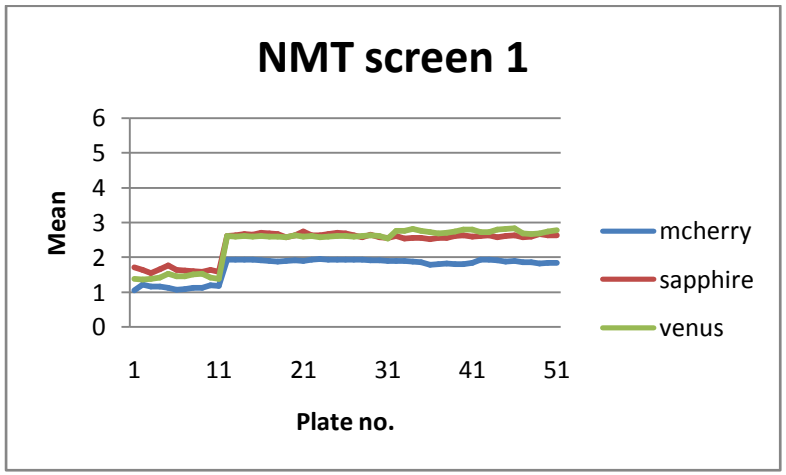


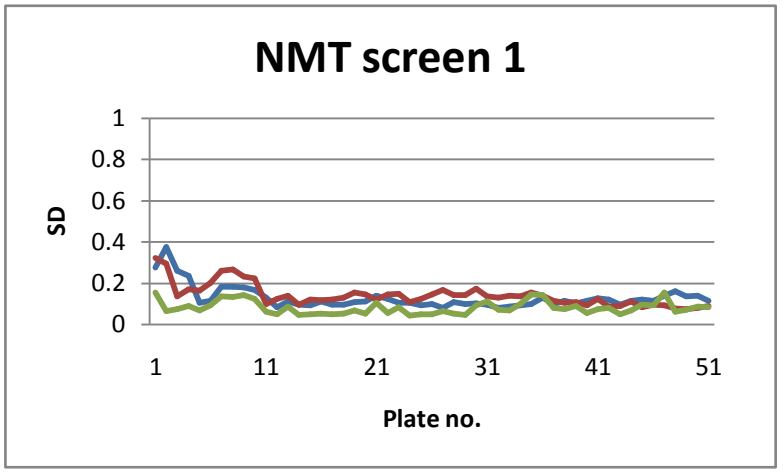


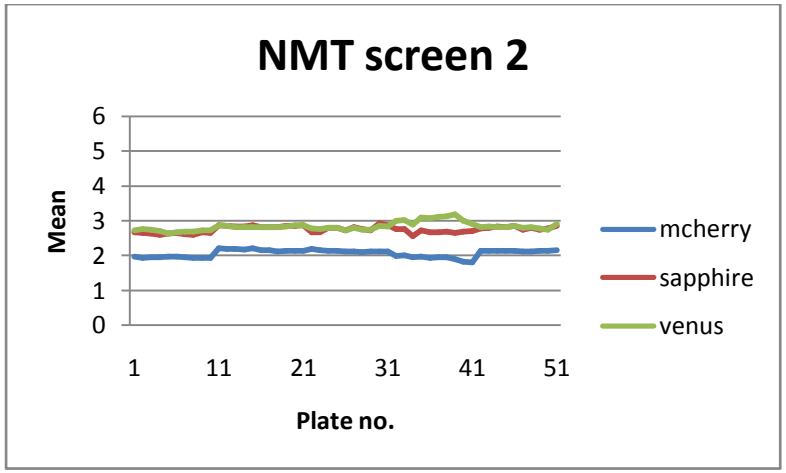
Note: Plate 15 of TS8 had DT mean of 1.8×10^{5} and SD of 1.3×10^{6}

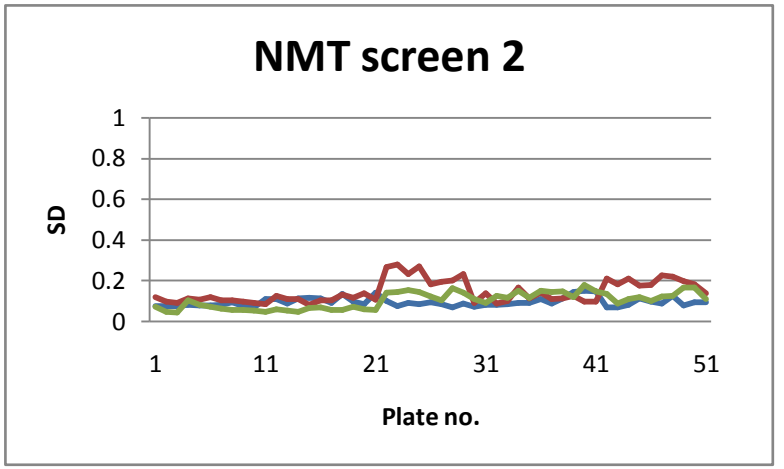

TS9 screen
Mean
8
7
6
5
4
3
2
1
0
1
11
21
31
41
51
Plate no.
mcherry
sapphire
venus

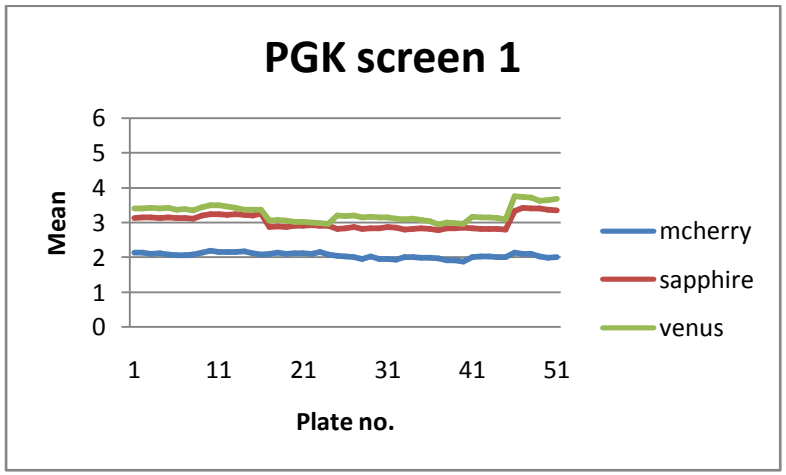

TS9 screen
SD
1
0.8
0.6
0.4
0.2
0
1
11
21
31
41
Plate no.

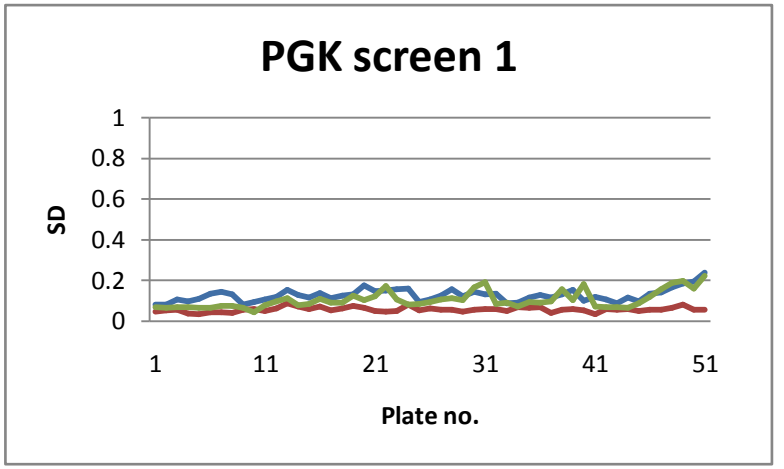

NMT screen 1
Mean
6
5
4
3
2
1
0
1
11
21
31
41
51
Plate no.
mcherry
sapphire
venus

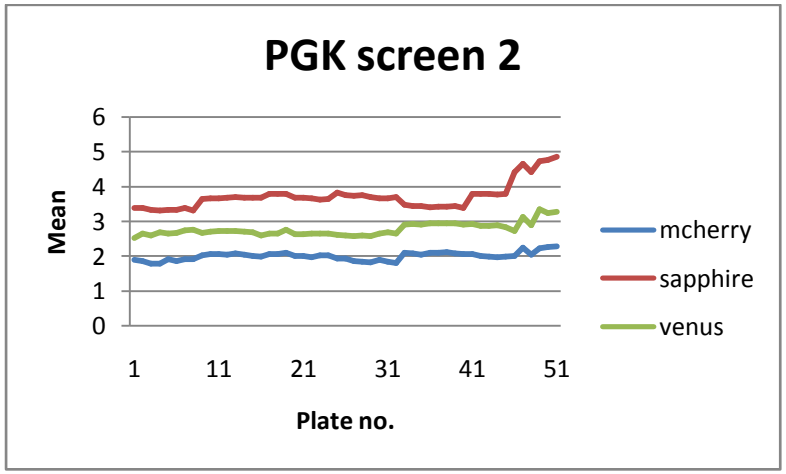

NMT screen 1
SD
1
0.8
0.6
0.4
0.2
0
1
11
21
31
41
51
Plate no.

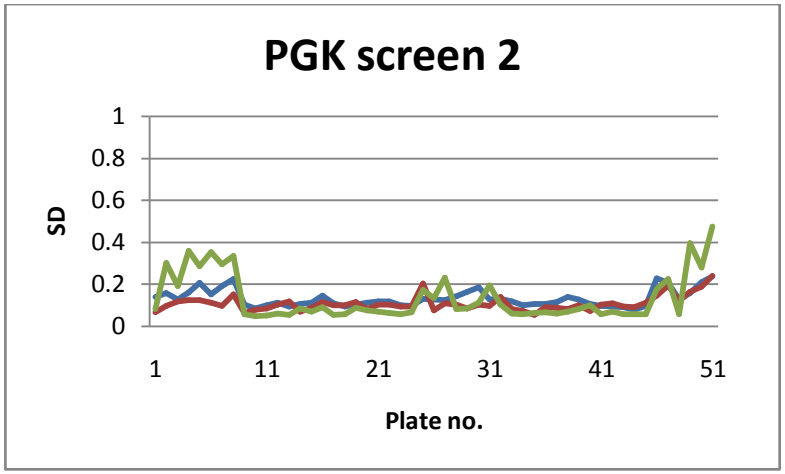

NMT screen 2
Mean
6
5
4
3
2
1
0
1
11
21
31
41
51
Plate no.
mcherry
sapphire
venus


NMT screen 2
SD
1
0.8
0.6
0.4
0.2
0
1
11
21
31
41
51
Plate no.


PGK screen 1
Mean
6
5
4
3
2
1
0
1
11
21
31
41
51
Plate no.
mcherry
sapphire
venus


PGK screen 1
SD
1
0.8
0.6
0.4
0.2
0
1
11
21
31
41
51
Plate no.


PGK screen 2
Mean
6
5
4
3
2
1
0
1
11
21
31
41
51
Plate no.
mcherry
sapphire
venus


PGK screen 2
SD
1
0.8
0.6
0.4
0.2
0
1
11
21
31
41
51
Plate no.

A.6: List of confirmation screens

Confirmation screens were generally run using 8 replicates at five concentrations (1, 2.5, 5, 10, 20 μm), or 4 replicates at a lower range of five concentrations (0.5, 1, 2.5, 5, 10 μm).

Assay	File ID	Concentration range, μm	Compounds		Replicates
			May	JHCCL	
TS3	Dec2010TS3cherrypick.csv	1 - 20	27*	-	8
		2, 10, 50			3
	CS_63_2_10_20110421150005.csv	1 - 20	53	5	8
	CS_63_3_32_20111209133124.csv	0.5 - 10	15*	4	4
	CS_63_7_45_20120417093418.csv	0.5 - 10	2**+5	14	4
	CS_63_12_61_20120531103405.csv	0.5 - 10	29**	2	4
TS4	CS_64_4_13_20110521153113.csv	1 - 20	54	8	8
	CS_64_5_23_20111028073121.csv	0.5 - 10	36*	-	4
	CS_64_6_31_20111209113125.csv	0.5 - 10	1*	3	4
TS5	CS_71_2_14_20110528141537.csv	1 - 20	53***	5	8
	CS_71_3_30_20111208133125.csv	0.5 - 10	10*	3	4
	CS_71_5_40_20120205070253.csv	0.5 - 10	0***	7	4
	CS_71_6_46_20120417093526.csv	0.5 - 10	5**+5	15	4
TS6	TS6cherrypick.csv	1 - 20	12**+17	16	8
	CS_77_3_6_20110325115514.csv	1 - 20	96	-	8
	CS_77_4_7_20110404123211.csv	1 - 20	102	-	8
	CS_77_5_9_20110411115111.csv	1 - 20	86	-	8
	CS_77_7_16_20110529120129.csv	1 - 20	63	-	8
	CS_77_8_24_20111028120222.csv	0.5 - 10	20*	-	4
	CS_77_10_37_20111212093549.csv	0.5 - 10	17**	-	4
	CS_77_11_41_20120205063144.csv	0.5 - 10	0***	7	4
	CS_77_14_51_20120427104149.csv	0.5 - 10	4**+6	15	4
	CS_77_15_62_20120607090640.csv	0.5 - 10	29**	2	4
TS7	CS_80_3_22_20110714113111.csv	0.5 - 10	37***+4	10	4
	CS_80_4_29_20111208110223.csv	0.5 - 10	10*	3	4
	CS_80_6_58_20120531103231.csv	0.5 - 10	30**+6	8	4
TS9	CS_85_2_38_20120127090445.csv	0.5 - 10	43*+1	3	4
	CS_85_3_63_20120615100158.csv	0.5 - 10	7*+1	3	4
PGK1	CS_72_2_17_20110603120112.csv	0.5 - 10	44	15	8
	CS_72_4_35_20111210133124.csv	0.5 - 10	21*	2	4
PGK2	CS_74_2_18_20110605120113.csv	0.5 - 10	50	12	4
	CS_74_4_36_20111212100237.csv	0.5 - 10	21*	1	4
NMT1	CS_78_2_19_20110609210111.csv	0.5 - 10	38+5	18	4
	CS_78_4_33_20111210110222.csv	0.5 - 10	9*	2	4
	CS_78_6_54_20120427104757.csv	0.5 - 10	4**+6	15	4
NMT2	CS_79_2_20_20110611133112.csv	0.5 - 10	38+1	22	4
	CS_79_6_53_20120427104440.csv	0.5 - 10	4**+6	15	4

* plus five positive controls

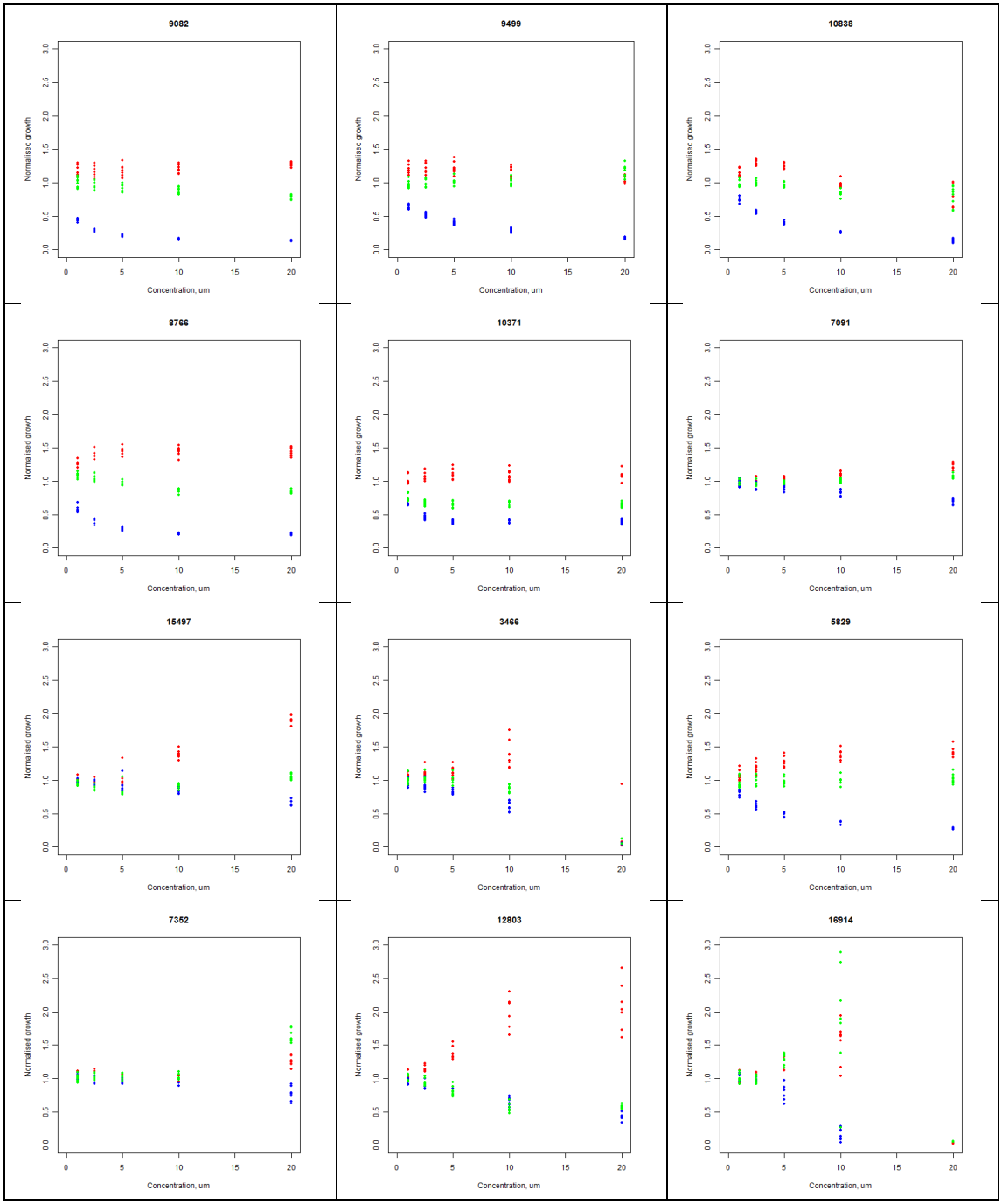
** plus three positive controls

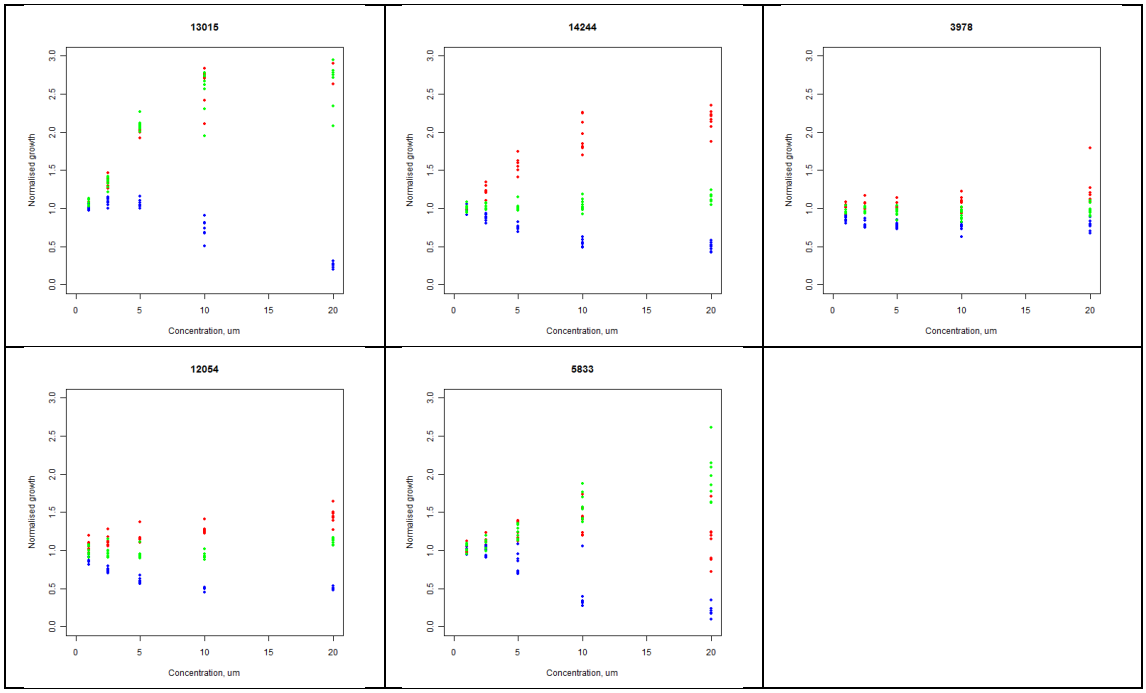
*** plus two positive controls

A.7 Confirmation curve examples

These curves are for 17 of the 20 strong PvDHFR candidates in Section 4.3.2. They are from TS6 CS_77_7_.

red=HsDHFR, blue=PvDHFR, green=PfDHFR.




13015
14244
3978
12054
5833
Normalised growth
Concentration, um

A.8 Confirmation: active Maybridge compounds

TS3 DHFR assay

	Eve ID	Maybridge ID	HsDHFR	PvDHFR	PfRdhfr	Pv active	PfR active	Toxic
TS3_Dec2010	510		0	74	0	Yes		No
TS3_Dec2010	516		0	44	0	Yes		No
TS3_Dec2010	978		0	80	0	Yes		No
TS3_63_7	978		1	40	0	Yes		No
TS3_63_12	978		0	40	0	Yes		No
TS3_63_2	3143	hfAW 00846	32	68	22	Yes	Yes	Possibly
TS3_63_2	3390	hfBTB 01480	0	52	0	Yes		No
TS3_Dec2010	3466	hfBTB 02152	16	30	8	Yes		Possibly
TS3_63_2	3474	hfBTB 02216	15	54	8	Yes		Possibly
TS3_63_2	3548	hfBTB 02678	0	37	0	Yes		No
TS3_Dec2010	3892	hfBTB 05160	41	43	40	Yes	Yes	Possibly
TS3_63_2	3947	hfBTB 05522	16	78	6	Yes		Possibly
TS3_Dec2010	3951	hfBTB 05541	34	61	51	Yes	Yes	Possibly
TS3_Dec2010	3978	hfBTB 05727	0	32	0	Yes		No
TS3_63_2	4305	hfBTB 08264	0	60	0	Yes		No
TS3_Dec2010	4848	hfBTB 13456	6	16	12	Yes	Yes	No
TS3_63_2	5678	hfCD 06694	13	16	7	Yes		Possibly
TS3_63_2	5809	hfCD 08381	18	42	54	Yes	Yes	Possibly
TS3_63_2	5822	hfCD 08497	0	60	0	Yes		No
TS3_Dec2010	5829	hfCD 08585	0	66	0	Yes		No
TS3_63_2	5868	hfCD 08965	0	56	0	Yes		No
TS3_63_2	6178	hfCD 11546	0	24	0	Yes		No
TS3_63_2	6254	hfDFP 00054	15	15	8	Yes		Possibly
TS3_Dec2010	6310	hfDP 00458	22	16	36	Yes	Yes	Possibly
TS3_63_2	7652	hfHTS 03328	0	56	0	Yes		No
TS3_63_2	8366	hfHTS 07614	0	72	0	Yes		No
TS3_Dec2010	8751	hfHTS 09910	0	56	0	Yes		No
TS3_63_2	9048	hfHTS 11969	0	42	0	Yes		No
TS3_Dec2010	9081	hfHTS 12148	0	80	0	Yes		No
TS3_Dec2010	9082	hfHTS 12152	7	80	4	Yes		No
TS3_63_2	9186	hfHTS 12551	0	80	0	Yes		No
TS3_Dec2010	9444	hfJFD 00261	7	76	1	Yes		No
TS3_Dec2010	9499	hfJFD 00787	13	78	6	Yes		Possibly
TS3_63_3	9504	hfJFD 00823	6	6	30		Yes	No
TS3_63_2	9525	hfJFD 00979	0	74	0	Yes		No
TS3_63_2	9557	hfJFD 01295	16	38	16	Yes	Yes	Possibly
TS3_63_2	9843	hfJFD 03375	0	74	0	Yes		No
TS3_Dec2010	10371	hfKM 03205	26	78	66	Yes	Yes	Possibly
TS3_63_2	10693	hfKM 05465	0	32	0	Yes		No
TS3_63_2	10879	hfKM 06831	2	32	2	Yes		No

	Eve ID	Maybridge ID	HsDHFR	PvDHFR	PfRdhfr	Pv active	PfR active	Toxic
TS3_63_2	11635	hfMWP 00601	0	50	0	Yes		No
TS3_Dec2010	11636	hfMWP 00602	0	22	0	Yes		No
TS3_63_2	11706	hfMWP 01127	0	78	0	Yes		No
TS3_Dec2010	11783	hfNRB 00102	16	32	16	Yes	Yes	Possibly
TS3_63_2	12037	hfNRB 03604	0	26	0	Yes		No
TS3_Dec2010	12054	hfNRB 03723	0	30	0	Yes		No
TS3_63_2	12100	hfNRB 04269	0	42	0	Yes		No
TS3_63_2	12255	hfPD 00426	0	0	10		Yes	No
TS3_63_2	12376	hfPHG 00991	0	44	0	Yes		No
TS3_63_2	12588	hfRDR 02635	0	66	0	Yes		No
TS3_63_2	12830	hfRF 02175	24	70	16	Yes	Yes	Possibly
TS3_Dec2010	13162	hfRH 00731	32	10	16	Yes	Yes	Possibly
TS3_63_2	14021	hfRJF 00951	0	80	0	Yes		No
TS3_63_2	14144	hfS 01961	0	52	0	Yes		No
TS3_63_2	14182	hfS 03874	0	66	0	Yes		No
TS3_63_2	14246	hfS 05379	0	54	0	Yes		No
TS3_Dec2010	14352	hfS 10015	28	8	28		Yes	Possibly
TS3_63_2	14447	hfS 12623	24	56	20	Yes	Yes	Possibly
TS3_63_2	14453	hfS 12745	0	38	32	Yes	Yes	No
TS3_63_2	14576	hfS 14685	7	46	8	Yes		No
TS3_63_2	15107	hfSEW 01466	0	54	0	Yes		No
TS3_63_2	15210	hfSEW 02156	0	54	10	Yes	Yes	No
TS3_Dec2010	15254	hfSEW 02484	24	40	24	Yes	Yes	Possibly
TS3_63_2	15667	hfSEW 05115	0	54	0	Yes		No
TS3_Dec2010	16020	hfSP 00278	0	80	0	Yes		No
TS3_63_12	16169	hfSP 01458	8	9	8	Yes	Yes	Possibly
TS3_Dec2010	16213	hfSPB 00471	28	32	29	Yes	Yes	Possibly
TS3_63_2	16219	hfSPB 00514	0	74	0	Yes		No
TS3_63_2	16490	hfSPB 02620	30	80	20	Yes	Yes	Possibly
TS3_63_2	16499	hfSPB 02669	0	54	0	Yes		No
TS3_63_2	16536	hfSPB 02854	0	64	0	Yes		No
TS3_Dec2010	16718	hfSPB 04137	26	32	26	Yes	Yes	Possibly
TS3_63_2	16756	hfSPB 04438	16	24	11	Yes	Yes	Possibly
TS3_Dec2010	16830	hfSPB 05131	14	66	9	Yes		Possibly
TS3_Dec2010	17006	hfSPB 06520	24	80	23	Yes	Yes	Possibly
TS3_63_2	17125	hfSPB 07412	0	30	0	Yes		No
TS3_63_2	17131	hfSPB 07441	0	40	0	Yes		No
TS3_63_2	17132	hfSPB 07445	0	66	0	Yes		No
TS3_Dec2010	17226	hfSPB 08252	3	46	0	Yes		No
TS3_63_2	17302	hfTL 00165	0	26	0	Yes		No

TS4 DHFR assay

	Eve ID	Maybridge ID	HsDHFR	TbDHFR	SmDHFR	Tb active	Sm active	Toxic
TS4_64_6	510		0	8	0	Yes		No
TS4_64_4	3258	hfBR 00082	23	22	20	Yes	Yes	Possibly
TS4_64_5	3259	hfBR 00086	8	8	8	Yes	Yes	Possibly
TS4_64_5	3626	hfBTB 03261	13	5	3			Possibly
TS4_64_5	3951	hfBTB 05541	11	24	19	Yes	Yes	Possibly
TS4_64_5	3978	hfBTB 05727	17	0	1			Possibly
TS4_64_4	4105	hfBTB 06669	6	20	2	Yes		No
TS4_64_4	4321	hfBTB 08347	30	56	62	Yes	Yes	Possibly
TS4_64_4	5173	hfCD 00513	0	16	0	Yes		No
TS4_64_4	5422	hfCD 03421	29	44	28	Yes	Yes	Possibly
TS4_64_4	5499	hfCD 04455	0	8	12		Yes	No
TS4_64_4	5506	hfCD 04510	20	23	15	Yes	Yes	Possibly
TS4_64_4	5833	hfCD 08635	33	42	29	Yes	Yes	Possibly
TS4_64_4	6247	hfDFP 00003	25	31	23	Yes	Yes	Possibly
TS4_64_4	6461	hfDP 01920	12	13	10	Yes	Yes	Possibly
TS4_64_5	6480	hfDSHS 00075	15	11	9	Yes	Yes	Possibly
TS4_64_4	6875	hfGK 03162	31	32	26	Yes	Yes	Possibly
TS4_64_4	7386	hfHTS 01930	14	23	17	Yes	Yes	Possibly
TS4_64_4	7512	hfHTS 02571	0	26	0	Yes		No
TS4_64_4	8440	hfHTS 08202	23	34	23	Yes	Yes	Possibly
TS4_64_4	9048	hfHTS 11969	0	16	0	Yes		No
TS4_64_5	9504	hfJFD 00823	0	8	12	Yes	Yes	No
TS4_64_4	9621	hfJFD 01902	13	9	9			Possibly
TS4_64_4	10041	hfKM 00407	10	16	11	Yes	Yes	Possibly
TS4_64_4	10063	hfKM 00585	4	24	2	Yes		No
TS4_64_4	10403	hfKM 03417	7	20	5	Yes		No
TS4_64_4	10537	hfKM 04403	0	6	16		Yes	No
TS4_64_4	10764	hfKM 06044	25	31	23	Yes	Yes	Possibly
TS4_64_5	10878	hfKM 06828	14	14	14	Yes	Yes	Possibly
TS4_64_5	10885	hfKM 06897	9	12	8	Yes	Yes	Possibly
TS4_64_4	11044	hfKM 08103	9	15	9	Yes		No
TS4_64_5	11133	hfKM 08617	16	16	13	Yes	Yes	Possibly
TS4_64_4	11164	hfKM 08832	32	16	9	Yes		Possibly
TS4_64_5	11250	hfKM 09319	9	6	5			Possibly
TS4_64_4	11614	hfMWP 00404	17	22	16	Yes	Yes	Possibly
TS4_64_5	11822	hfNRB 00390	15	11	10	Yes	Yes	Possibly
TS4_64_4	12000	hfNRB 02920	0	20	26	Yes	Yes	No
TS4_64_4	12007	hfNRB 03047	14	18	9	Yes		Possibly
TS4_64_4	12234	hfPD 00323	3	16	1	Yes		No
TS4_64_4	12251	hfPD 00407	26	29	23	Yes	Yes	Possibly
TS4_64_4	12252	hfPD 00408	26	23	17	Yes	Yes	Possibly

	Eve ID	Maybridge ID	HsDHFR	TbDHFR	SmDHFR	Tb active	Sm active	Toxic
TS4_64_5	12803	hfRF 01744	5	16	16	Yes	Yes	No
TS4_64_4	12868	hfRF 02895	10	15	12	Yes	Yes	Possibly
TS4_64_4	14117	hfS 00540	32	31	25	Yes	Yes	Possibly
TS4_64_4	14129	hfS 01394	2	18	1	Yes		No
TS4_64_4	14244	hfS 05363	6	35	17	Yes	Yes	No
TS4_64_4	14447	hfS 12623	25	34	23	Yes	Yes	Possibly
TS4_64_4	14528	hfS 14125	0	16	0	Yes		No
TS4_64_4	14874	hfSCR 01008	30	32	28	Yes	Yes	Possibly
TS4_64_4	15309	hfSEW 02839	19	21	21	Yes	Yes	Possibly
TS4_64_4	15360	hfSEW 03401	4	44	2	Yes		No
TS4_64_4	15393	hfSEW 03591	12	14	12	Yes	Yes	Possibly
TS4_64_5	16217	hfSPB 00506	19	11	9	Yes	Yes	Possibly
TS4_64_5	16236	hfSPB 00625	10	8	4	Yes		Possibly
TS4_64_4	16599	hfSPB 03238	14	18	14	Yes	Yes	Possibly
TS4_64_5	16718	hfSPB 04137	9	8	6	Yes		Possibly
TS4_64_4	16756	hfSPB 04438	19	28	17	Yes	Yes	Possibly

TS5 DHFR assay

	Eve ID	Maybridge ID	PfRdhfr	HsDHFR	TcDHFR	PfR active	Tc active	Toxic
TS5_71_2	510		0	0	74		Yes	No
TS5_71_2	3040	hfAW 00264	12	16	16	Yes	Yes	Possibly
TS5_71_2	3594	hfBTB 02990	14	8	16	Yes	Yes	No
TS5_71_2	3892	hfBTB 05160	0	17	42		Yes	Possibly
TS5_71_2	3978	hfBTB 05727	27	31	43	Yes	Yes	Possibly
TS5_71_2	4321	hfBTB 08347	54	25	58	Yes	Yes	Possibly
TS5_71_3	4321	hfBTB 08347	16	10	8	Yes	Yes	Possibly
TS5_71_2	4606	hfBTB 10539	38	0	0	Yes		No
TS5_71_2	4655	hfBTB 11167	38	15	40	Yes	Yes	Possibly
TS5_71_2	4699	hfBTB 11765	30	18	24	Yes	Yes	Possibly
TS5_71_2	4880	hfBTB 13766	42	28	43	Yes	Yes	Possibly
TS5_71_2	4939	hfBTB 14154	46	9	16	Yes	Yes	No
TS5_71_2	5322	hfCD 02264	35	24	37	Yes	Yes	Possibly
TS5_71_2	5397	hfCD 03092	38	0	18	Yes	Yes	No
TS5_71_2	5422	hfCD 03421	27	24	36	Yes	Yes	Possibly
TS5_71_2	5499	hfCD 04455	0	0	42		Yes	No
TS5_71_2	6083	hfCD 10740	0	0	12		Yes	No
TS5_71_2	6480	hfDSHS 00075	30	32	32	Yes	Yes	Possibly
TS5_71_2	6600	hfEN 00275	44	0	0	Yes		No
TS5_71_2	6923	hfGK 03531	30	0	13	Yes	Yes	No
TS5_71_2	7107	hfHTS 00290	0	16	18		Yes	Possibly

	Eve ID	Maybridge ID	PfRdhfr	HsDHFR	TcRdhfr	PfR active	Tc active	Toxic
TS5_71_2	7145	hfHTS 00480	58	0	0	Yes		No
TS5_71_2	8286	hfHTS 07134	48	6	46	Yes	Yes	No
TS5_71_2	8838	hfHTS 10391	36	0	0	Yes		No
TS5_71_2	9444	hfJFD 00261	7	0	12		Yes	No
TS5_71_2	9499	hfJFD 00787	10	10	22	Yes	Yes	Possibly
TS5_71_3	9504	hfJFD 00823	0	8	10		Yes	Possibly
TS5_71_2	9976	hfJP 00899	44	14	16	Yes	Yes	Possibly
TS5_71_2	11133	hfKM 08617	27	30	47	Yes	Yes	Possibly
TS5_71_2	11164	hfKM 08832	31	3	10	Yes	Yes	No
TS5_71_2	11250	hfKM 09319	34	23	34	Yes	Yes	Possibly
TS5_71_2	12054	hfNRB 03723	7	6	10		Yes	No
TS5_71_2	12395	hfRB 00158	24	24	32	Yes	Yes	Possibly
TS5_71_2	12706	hfRF 00305	58	0	0	Yes		No
TS5_71_3	12803	hfRF 01744	4	10	10		Yes	Possibly
TS5_71_2	12830	hfRF 02175	7	7	37		Yes	No
TS5_71_2	12868	hfRF 02895	9	7	22		Yes	No
TS5_71_2	12923	hfRF 03874	48	16	1	Yes		Possibly
TS5_71_2	13162	hfRH 00731	30	6	38	Yes	Yes	No
TS5_71_2	14244	hfS 05363	1	17	11		Yes	Possibly
TS5_71_2	14342	hfS 09668	56	8	17	Yes	Yes	No
TS5_71_2	15179	hfSEW 01952	44	0	10	Yes	Yes	No
TS5_71_2	15182	hfSEW 01975	46	0	17	Yes	Yes	No
TS5_71_2	15469	hfSEW 03989	46	16	15	Yes	Yes	Possibly
TS5_71_2	16207	hfSPB 00432	5	8	24		Yes	No
TS5_71_2	16213	hfSPB 00471	29	30	40	Yes	Yes	Possibly
TS5_71_2	16222	hfSPB 00523	58	2	33	Yes	Yes	No
TS5_71_2	16250	hfSPB 00828	22	0	0	Yes		No
TS5_71_2	16252	hfSPB 00832	36	0	9	Yes		No
TS5_71_2	16486	hfSPB 02590	14	0	0	Yes		No
TS5_71_2	16490	hfSPB 02620	23	27	42	Yes	Yes	Possibly
TS5_71_2	16830	hfSPB 05131	12	18	34	Yes	Yes	Possibly
TS5_71_2	16914	hfSPB 05912	33	33	34	Yes	Yes	Possibly

TS6 DHFR assay

	Eve ID	Maybridge ID	HsDHFR	PvDHFR	PfDHFR	Pv active	Pf active	Toxic
TS6_c'pick	510		0	76	0	Yes		No
TS6_c'pick	516		0	58	14	Yes	Yes	No
TS6_c'pick	978		0	80	80	Yes	Yes	No
TS6_77_8	978		0	24	2	Yes		No
TS6_77_10	978		0	18	0	Yes		No
TS6_77_11	978		0	30	4	Yes		No
TS6_77_14	978		0	40	18	Yes	Yes	No
TS6_77_15	978		0	40	14	Yes	Yes	No
TS6_c'pick	3143	hfAW 00846	21	66	22	Yes	Yes	Possibly
TS6_77_8	3259	hfBR 00086	8	0	7			Possibly
TS6_77_7	3466	hfBTB 02152	16	36	18	Yes	Yes	Possibly
TS6_77_16	3466	hfBTB 02152	8	8	8	Yes	Yes	Possibly
TS6_c'pick	3548	hfBTB 02678	0	18	0	Yes		No
TS6_77_7	3594	hfBTB 02990	8	30	16	Yes	Yes	No
TS6_77_7	3740	hfBTB 04150	0	22	52	Yes	Yes	No
TS6_77_5	3947	hfBTB 05522	17	6	17		Yes	Possibly
TS6_77_10	3951	hfBTB 05541	18	8	9	Yes	Yes	Possibly
TS6_77_16	3951	hfBTB 05541	13	5	7			Possibly
TS6_77_7	3978	hfBTB 05727	14	56	15	Yes	Yes	Possibly
TS6_77_16	3978	hfBTB 05727	8	7	7			Possibly
TS6_77_7	4584	hfBTB 10320	16	25	19	Yes	Yes	Possibly
TS6_77_7	4662	hfBTB 11323	9	80	14	Yes	Yes	No
TS6_77_7	4769	hfBTB 12704	0	0	26		Yes	No
TS6_77_3	4867	hfBTB 13657	14	46	8	Yes		Possibly
TS6_77_10	4879	hfBTB 13762	12	1	0			Possibly
TS6_77_10	4939	hfBTB 14154	8	0	0			Possibly
TS6_77_7	5422	hfCD 03421	29	44	23	Yes	Yes	Possibly
TS6_77_10	5422	hfCD 03421	0	0	16		Yes	No
TS6_77_8	5499	hfCD 04455	0	0	12		Yes	No
TS6_c'pick	5678	hfCD 06694	16	18	16	Yes	Yes	Possibly
TS6_77_3	5678	hfCD 06694	16	24	16	Yes	Yes	Possibly
TS6_77_16	5678	hfCD 06694	11	12	11	Yes	Yes	Possibly
TS6_77_7	5829	hfCD 08585	0	72	0	Yes		No
TS6_77_7	5833	hfCD 08635	30	36	21	Yes	Yes	Possibly
TS6_77_7	5884	hfCD 09120	0	8	42		Yes	No
TS6_77_3	6049	hfCD 10467	12	0	18		Yes	Possibly
TS6_77_8	6480	hfDSHS 00075	12	9	12	Yes	Yes	Possibly
TS6_77_7	7091	hfHTS 00223	0	22	0	Yes		No
TS6_77_7	7283	hfHTS 01369	0	38	0	Yes		No
TS6_77_7	7352	hfHTS 01717	10	13	8	Yes		Possibly
TS6_77_7	7477	hfHTS 02381	0	22	26	Yes	Yes	No

	Eve ID	Maybridge ID	HsDHFR	PvDHFR	PfDHFR	Pv active	Pf active	Toxic
TS6_77_7	8028	hfHTS 05561	0	14	32	Yes	Yes	No
TS6_77_7	8751	hfHTS 09910	0	58	0	Yes		No
TS6_77_7	8766	hfHTS 09988	0	80	2	Yes		No
TS6_77_15	8853	hfHTS 10441	0	0	8		Yes	No
TS6_77_7	9082	hfHTS 12152	2	80	14	Yes	Yes	No
TS6_c'pick	9156	hfHTS 12446	2	0	18		Yes	No
TS6_77_4	9156	hfHTS 12446	0	0	14		Yes	No
TS6_77_7	9186	hfHTS 12551	0	80	0	Yes		No
TS6_77_4	9286	hfHTS 12958	20	0	8			Possibly
TS6_77_5	9348	hfHTS 13283	8	2	14		Yes	No
TS6_77_4	9479	hfJFD 00597	0	36	0	Yes		No
TS6_77_7	9499	hfJFD 00787	8	80	10	Yes	Yes	No
TS6_77_16	9499	hfJFD 00787	10	10	10	Yes	Yes	Possibly
TS6_77_7	9504	hfJFD 00823	20	44	60	Yes	Yes	Possibly
TS6_77_8	9504	hfJFD 00823	0	8	28	Yes	Yes	No
TS6_77_7	9525	hfJFD 00979	0	80	0	Yes		No
TS6_77_7	9559	hfJFD 01325	0	68	0	Yes		No
TS6_77_7	9762	hfJFD 02848	0	30	16	Yes	Yes	No
TS6_77_7	9843	hfJFD 03375	0	78	18	Yes	Yes	No
TS6_77_7	9853	hfJFD 03520	8	8	11		Yes	No
TS6_77_7	9892	hfJFD 03913	13	26	14	Yes	Yes	Possibly
TS6_77_7	10371	hfKM 03205	0	80	76	Yes	Yes	No
TS6_77_7	10655	hfKM 05251	0	62	0	Yes		No
TS6_77_4	10684	hfKM 05413	0	24	6	Yes		No
TS6_77_7	10708	hfKM 05576	12	74	52	Yes	Yes	Possibly
TS6_77_7	10709	hfKM 05590	0	60	52	Yes	Yes	No
TS6_77_7	10737	hfKM 05853	0	30	0	Yes		No
TS6_77_7	10838	hfKM 06626	21	80	31	Yes	Yes	Possibly
TS6_77_3	10874	hfKM 06811	15	11	8	Yes		Possibly
TS6_77_7	10878	hfKM 06828	46	56	43	Yes	Yes	Possibly
TS6_c'pick	10879	hfKM 06831	0	30	1	Yes		No
TS6_77_7	10974	hfKM 07711	0	78	0	Yes		No
TS6_77_7	11133	hfKM 08617	30	46	19	Yes	Yes	Possibly
TS6_77_8	11250	hfKM 09319	12	2	5			Possibly
TS6_77_7	11663	hfMWP 00824	0	44	3	Yes		No
TS6_c'pick	11706	hfMWP 01127	0	74	60	Yes	Yes	No
TS6_77_16	11783	hfNRB 00102	13	12	12	Yes	Yes	Possibly
TS6_77_8	11822	hfNRB 00390	8	4	8		Yes	Possibly
TS6_77_3	12017	hfNRB 03257	0	16	0	Yes		No
TS6_77_7	12054	hfNRB 03723	7	64	9	Yes		No
TS6_77_7	12111	hfNRB 04514	0	78	0	Yes		No
TS6_77_7	12382	hfRB 00019	38	80	48	Yes	Yes	Possibly

	Eve ID	Maybridge ID	HsDHFR	PvDHFR	PfDHFR	Pv active	Pf active	Toxic
TS6_77_7	12588	hfRDR 02635	0	70	24	Yes	Yes	No
TS6_77_7	12803	hfRF 01744	16	36	42	Yes	Yes	Possibly
TS6_77_8	12803	hfRF 01744	4	4	16		Yes	No
TS6_c'pick	12830	hfRF 02175	19	22	33	Yes	Yes	Possibly
TS6_77_4	12830	hfRF 02175	23	60	36	Yes	Yes	Possibly
TS6_77_3	12913	hfRF 03548	10	80	17	Yes	Yes	Possibly
TS6_77_7	13003	hfRF 04999	0	29	28	Yes	Yes	No
TS6_77_7	13015	hfRF 05142	9	28	8	Yes		No
TS6_77_3	13485	hfRJC 00408	8	16	26	Yes	Yes	No
TS6_77_7	13664	hfRJC 02246	0	22	28	Yes	Yes	No
TS6_77_7	14021	hfRJF 00951	0	80	14	Yes	Yes	No
TS6_77_7	14238	hfS 05244	32	46	26	Yes	Yes	Possibly
TS6_77_7	14244	hfS 05363	14	48	19	Yes	Yes	Possibly
TS6_77_7	14246	hfS 05379	0	56	2	Yes		No
TS6_77_7	14329	hfS 08684	0	2	28		Yes	No
TS6_77_7	14575	hfS 14676	0	56	0	Yes		No
TS6_c'pick	14576	hfS 14685	0	30	0	Yes		No
TS6_77_4	14952	hfSEW 00102	18	62	8	Yes		Possibly
TS6_77_16	15254	hfSEW 02484	8	4	6			Possibly
TS6_77_7	15497	hfSEW 04168	8	20	21	Yes	Yes	No
TS6_77_7	15761	hfSEW 05596	0	70	0	Yes		No
TS6_77_7	16169	hfSP 01458	39	46	39	Yes	Yes	Possibly
TS6_77_15	16169	hfSP 01458	10	14	12	Yes	Yes	Possibly
TS6_77_4	16172	hfSP 01461	0	12	4	Yes		No
TS6_77_7	16211	hfSPB 00468	0	30	60	Yes	Yes	No
TS6_77_7	16212	hfSPB 00470	10	34	62	Yes	Yes	Possibly
TS6_77_16	16213	hfSPB 00471	16	16	16	Yes	Yes	Possibly
TS6_77_8	16213	hfSPB 00471	9	1	5			Possibly
TS6_77_10	16236	hfSPB 00625	8	1	5			Possibly
TS6_77_4	16349	hfSPB 01622	0	32	0	Yes		No
TS6_77_4	16350	hfSPB 01624	0	32	0	Yes		No
TS6_c'pick	16490	hfSPB 02620	19	78	22	Yes	Yes	Possibly
TS6_77_4	16490	hfSPB 02620	18	78	12	Yes	Yes	Possibly
TS6_77_16	16490	hfSPB 02620	14	17	13	Yes	Yes	Possibly
TS6_c'pick	16499	hfSPB 02669	0	62	0	Yes		No
TS6_77_4	16499	hfSPB 02669	0	62	0	Yes		No
TS6_77_4	16556	hfSPB 02947	0	0	18		Yes	No
TS6_77_7	16687	hfSPB 03954	0	55	3	Yes		No
TS6_77_16	16718	hfSPB 04137	8	8	8	Yes	Yes	Possibly
TS6_77_3	16830	hfSPB 05131	23	80	38	Yes	Yes	Possibly
TS6_77_7	16914	hfSPB 05912	33	38	26	Yes	Yes	Possibly
TS6_77_5	17095	hfSPB 07211	32	24	32	Yes	Yes	Possibly
TS6_77_4	17097	hfSPB 07218	0	56	0	Yes		No

	Eve ID	Maybridge ID	HsDHFR	PvDHFR	PfDHFR	Pv active	Pf active	Toxic
TS6_77_3	17167	hfSPB 07894	31	80	54	Yes	Yes	Possibly
TS6_77_4	17170	hfSPB 07935	30	80	49	Yes	Yes	Possibly
TS6_77_5	17217	hfSPB 08198	45	47	46	Yes	Yes	Possibly
TS6_77_4	17226	hfSPB 08252	27	80	32	Yes	Yes	Possibly
TS6_77_16	17226	hfSPB 08252	10	12	12	Yes	Yes	Possibly
TS6_77_7	17302	hfTL 00165	0	54	0	Yes		No
TS6_77_7	17329	hfXAX 00025	5	80	13	Yes	Yes	No
TS6_77_16	18008	SMSPB06943SC	8	10	10	Yes	Yes	Possibly
TS6_c'pick	18010	SMHTS12151SC	13	80	17	Yes	Yes	Possibly
TS6_c'pick	18011	SMHTS12147SC	10	80	15	Yes	Yes	Possibly
TS6_c'pick	18012	SMHTS07613SC	0	58	0	Yes		No
TS6_c'pick	18014	SMHTS12150SC	48	80	44	Yes	Yes	Possibly

TS7 DHFR assay

	Eve ID	Maybridge ID	HsDHFR	LmDHFR	PvRdhfr	Lm active	PvR active	Toxic
TS7_80_3	978		0	0	28		Yes	No
TS7_80_4	4321	hfBTB 08347	16	12	20	Yes	Yes	Possibly
TS7_80_6	4321	hfBTB 08347	12	10	14	Yes	Yes	Possibly
TS7_80_3	4435	hfBTB 09154	0	0	18		Yes	No
TS7_80_6	4939	hfBTB 14154	8	0	8		Yes	Possibly
TS7_80_6	5403	hfCD 03158	16	16	8	Yes	Yes	Possibly
TS7_80_3	5829	hfCD 08585	0	0	24		Yes	No
TS7_80_3	5833	hfCD 08635	12	8	12	Yes	Yes	Possibly
TS7_80_3	6173	hfCD 11507	4	0	34		Yes	No
TS7_80_6	6173	hfCD 11507	12	5	8		Yes	Possibly
TS7_80_3	6480	hfDSHS 00075	8	8	3	Yes		Possibly
TS7_80_3	6875	hfGK 03162	8	0	24		Yes	Possibly
TS7_80_3	8440	hfHTS 08202	3	0	12		Yes	No
TS7_80_3	9444	hfJFD 00261	0	0	22		Yes	No
TS7_80_3	9499	hfJFD 00787	0	0	26		Yes	No
TS7_80_6	9499	hfJFD 00787	9	9	3	Yes		Possibly
TS7_80_3	9504	hfJFD 00823	0	8	30	Yes	Yes	No
TS7_80_4	9504	hfJFD 00823	0	8	18	Yes	Yes	No
TS7_80_6	9504	hfJFD 00823	0	8	22	Yes	Yes	No
TS7_80_3	9525	hfJFD 00979	0	0	22		Yes	No
TS7_80_3	9876	hfJFD 03674	0	0	22		Yes	No
TS7_80_3	10371	hfKM 03205	2	16	22	Yes	Yes	No
TS7_80_3	10608	hfKM 04905	0	4	10		Yes	No
TS7_80_3	10885	hfKM 06897	1	1	9		Yes	No
TS7_80_3	11636	hfMWP 00602	0	0	26		Yes	No
TS7_80_3	11783	hfNRB 00102	8	8	16	Yes	Yes	Possibly

	Eve ID	Maybridge ID	HsDHFR	LmDHFR	PvRdhfr	Lm active	PvR active	Toxic
TS7_80_4	12803	hfRF 01744	6	16	0	Yes		No
TS7_80_6	13361	hfRH 01876	26	32	13	Yes	Yes	Possibly
TS7_80_3	13361	hfRH 01876	12	8	4	Yes		Possibly
TS7_80_3	14246	hfS 05379	0	0	22		Yes	No
TS7_80_3	14371	hfS 10607	18	22	3	Yes		Possibly
TS7_80_3	15254	hfSEW 02484	8	6	14		Yes	Possibly
TS7_80_6	15254	hfSEW 02484	9	4	16		Yes	Possibly
TS7_80_3	15667	hfSEW 05115	0	0	8		Yes	No
TS7_80_3	16211	hfSPB 00468	0	2	18		Yes	No
TS7_80_6	16211	hfSPB 00468	0	0	14		Yes	No
TS7_80_3	16212	hfSPB 00470	0	4	20		Yes	No
TS7_80_6	16212	hfSPB 00470	0	0	18		Yes	No
TS7_80_3	16217	hfSPB 00506	14	12	24	Yes	Yes	Possibly
TS7_80_3	16219	hfSPB 00514	0	0	10		Yes	No
TS7_80_3	16499	hfSPB 02669	0	0	16		Yes	No
TS7_80_6	16718	hfSPB 04137	14	14	8	Yes	Yes	Possibly
TS7_80_3	16718	hfSPB 04137	5	8	2	Yes		No
TS7_80_6	16724	hfSPB 04186	10	8	6	Yes		Possibly
TS7_80_3	16830	hfSPB 05131	2	1	26		Yes	No
TS7_80_3	17167	hfSPB 07894	4	2	28		Yes	No
TS7_80_3	17182	hfSPB 07993	0	0	16		Yes	No
TS7_80_3	17329	hfXAX 00025	0	0	36		Yes	No
TS7_80_3	18010	SMHTS12151SC	0	0	24		Yes	No
TS7_80_6	18010	SMHTS12151SC	5	8	0	Yes		No
TS7_80_6	18011	SMHTS12147SC	22	22	0	Yes		Possibly
TS7_80_3	18012	SMHTS07613SC	0	0	16		Yes	No
TS7_80_3	18014	SMHTS12150SC	0	0	24		Yes	No

NMT assay 1

	Eve ID	Maybridge ID	HsNMT	TbNMT	PvNMT	Tb active	Pv active	Toxic
NMT_78_2	3259	hfBR 00086	1	8	6	Yes		No
NMT_78_2	3320	hfBTB 00809	0	16	8	Yes	Yes	No
NMT_78_2	3378	hfBTB 01383	0	30	32	Yes	Yes	No
NMT_78_4	5422	hfCD 03421	0	2	30		Yes	No
NMT_78_2	5695	hfCD 06957	0	24	22	Yes	Yes	No
NMT_78_2	5699	hfCD 07074	0	12	0	Yes		No
NMT_78_2	5971	hfCD 09895	0	22	0	Yes		No
NMT_78_2	6254	hfDFP 00054	0	12	10	Yes	Yes	No
NMT_78_2	6353	hfDP 00892	3	12	12	Yes	Yes	No
NMT_78_4	6777	hfGK 01974	8	6	30		Yes	Possibly

	Eve ID	Maybridge ID	HsNMT	TbNMT	PvNMT	Tb active	Pv active	Toxic
NMT_78_2	8028	hfHTS 05561	0	32	30	Yes	Yes	No
NMT_78_6	9046	hfHTS 11966	18	0	18		Yes	Possibly
NMT_78_6	9080	hfHTS 12146	14	0	8		Yes	Possibly
NMT_78_6	9081	hfHTS 12148	0	12	0	Yes		No
NMT_78_2	9694	hfJFD 02390	0	26	10	Yes	Yes	No
NMT_78_4	9762	hfJFD 02848	14	14	20	Yes	Yes	Possibly
NMT_78_2	9877	hfJFD 03675	0	14	0	Yes		No
NMT_78_2	10708	hfKM 05576	0	10	3	Yes		No
NMT_78_4	10838	hfKM 06626	0	8	0	Yes		No
NMT_78_2	11133	hfKM 08617	5	28	26	Yes	Yes	No
NMT_78_2	12506	hfRDR 01446	6	34	36	Yes	Yes	No
NMT_78_2	12803	hfRF 01744	4	9	9	Yes	Yes	No
NMT_78_4	12803	hfRF 01744	0	8	4	Yes		No
NMT_78_2	12830	hfRF 02175	0	1	8		Yes	No
NMT_78_2	12868	hfRF 02895	0	10	8	Yes	Yes	No
NMT_78_2	13003	hfRF 04999	3	24	16	Yes	Yes	No
NMT_78_2	13664	hfRJC 02246	0	24	24	Yes	Yes	No
NMT_78_2	13687	hfRJC 02395	0	4	16		Yes	No
NMT_78_2	14238	hfS 05244	5	14	8	Yes	Yes	No
NMT_78_2	14345	hfS 09767	0	16	14	Yes	Yes	No
NMT_78_2	16211	hfSPB 00468	0	18	28	Yes	Yes	No
NMT_78_2	16212	hfSPB 00470	0	22	28	Yes	Yes	No
NMT_78_2	16217	hfSPB 00506	5	26	30	Yes	Yes	No
NMT_78_2	16718	hfSPB 04137	5	22	20	Yes	Yes	No
NMT_78_2	16914	hfSPB 05912	6	16	12	Yes	Yes	No
NMT_78_2	17196	hfSPB 08060	0	28	20	Yes	Yes	No
NMT_78_2	17399	hfXBX 00332	0	12	6	Yes		No
NMT_78_2	18004	SMSPB05423SC	0	22	0	Yes		No
NMT_78_2	18009	SMSPB05424SC	0	8	0	Yes		No
NMT_78_2	18010	SMHTS12151SC	0	24	12	Yes	Yes	No
NMT_78_6	18010	SMHTS12151SC	0	12	0	Yes		No
NMT_78_2	18011	SMHTS12147SC	0	36	26	Yes	Yes	No
NMT_78_6	18011	SMHTS12147SC	0	20	0	Yes		No
NMT_78_2	18014	SMHTS12150SC	0	14	0	Yes		No

NMT assay 2

	Eve ID	Maybridge ID	HsNMT	SmNMT	TcNMT	Sm active	Tc active	Toxic
NMT_79_2	3740	hfBTB 04150	0	10	22	Yes	Yes	No
NMT_79_2	3951	hfBTB 05541	22	0	0			Possibly
NMT_79_2	3999	hfBTB 05867	0	0	22		Yes	No
NMT_79_2	4495	hfBTB 09584	6	2	20		Yes	No
NMT_79_2	4496	hfBTB 09587	4	2	18		Yes	No
NMT_79_2	5499	hfCD 04455	0	0	12		Yes	No
NMT_79_2	5699	hfCD 07074	0	0	12		Yes	No
NMT_79_2	5809	hfCD 08381	16	24	22	Yes	Yes	Possibly
NMT_79_2	5833	hfCD 08635	4	7	8		Yes	No
NMT_79_2	7107	hfHTS 00290	8	0	8		Yes	Possibly
NMT_79_2	8028	hfHTS 05561	0	4	26		Yes	No
NMT_79_6	9046	hfHTS 11966	16	0	14		Yes	Possibly
NMT_79_6	9080	hfHTS 12146	12	0	8		Yes	Possibly
NMT_79_6	9082	hfHTS 12152	10	0	0			Possibly
NMT_79_2	9504	hfJFD 00823	8	20	34	Yes	Yes	Possibly
NMT_79_2	9588	hfJFD 01579	2	0	12		Yes	No
NMT_79_2	10121	hfKM 01046	0	0	12		Yes	No
NMT_79_2	10764	hfKM 06044	2	0	8		Yes	No
NMT_79_2	10838	hfKM 06626	4	8	6	Yes		No
NMT_79_2	11589	hfMWP 00123	8	6	16		Yes	Possibly
NMT_79_2	12005	hfNRB 03005	0	0	14		Yes	No
NMT_79_2	12630	hfRDR 03524	0	10	10	Yes	Yes	No
NMT_79_2	12803	hfRF 01744	6	0	8		Yes	No
NMT_79_2	12978	hfRF 04603	0	0	10		Yes	No
NMT_79_2	13059	hfRH 00058	0	0	8		Yes	No
NMT_79_2	13664	hfRJC 02246	0	16	26	Yes	Yes	No
NMT_79_2	14345	hfS 09767	0	4	14		Yes	No
NMT_79_2	14371	hfS 10607	14	20	16	Yes	Yes	Possibly
NMT_79_2	15497	hfSEW 04168	0	0	22		Yes	No
NMT_79_2	15656	hfSEW 04978	0	8	28	Yes	Yes	No
NMT_79_2	16211	hfSPB 00468	0	6	24		Yes	No
NMT_79_2	16212	hfSPB 00470	0	10	24	Yes	Yes	No
NMT_79_2	17006	hfSPB 06520	8	8	14	Yes	Yes	Possibly
NMT_79_2	18008	SMSPB06943SC	16	16	16	Yes	Yes	Possibly

PGK assay 1

	Eve ID	Maybridge ID	HsPGK	SmPGK	TcPGK	Sm active	Tc active	Toxic
PGK1_72_2	3259	hfBR 00086	24	22	32	Yes	Yes	Possibly
PGK1_72_2	3594	hfBTB 02990	14	16	16	Yes	Yes	Possibly
PGK1_72_2	3951	hfBTB 05541	22	27	62	Yes	Yes	Possibly
PGK1_72_2	4105	hfBTB 06669	0	14	0	Yes		No
PGK1_72_2	4655	hfBTB 11167	16	17	48	Yes	Yes	Possibly
PGK1_72_2	5403	hfCD 03158	33	37	25	Yes	Yes	Possibly
PGK1_72_2	5422	hfCD 03421	31	32	31	Yes	Yes	Possibly
PGK1_72_2	5595	hfCD 05564	44	20	31	Yes	Yes	Possibly
PGK1_72_2	5909	hfCD 09340	0	0	30		Yes	No
PGK1_72_2	6173	hfCD 11507	24	16	16	Yes	Yes	Possibly
PGK1_72_4	6777	hfGK 01974	1	0	16		Yes	No
PGK1_72_2	7259	hfHTS 01223	1	16	0	Yes		No
PGK1_72_2	8028	hfHTS 05561	0	0	42		Yes	No
PGK1_72_2	8029	hfHTS 05567	0	0	28		Yes	No
PGK1_72_2	9203	hfHTS 12635	32	32	32	Yes	Yes	Possibly
PGK1_72_2	9504	hfJFD 00823	0	16	16	Yes	Yes	No
PGK1_72_2	9912	hfJFD 03992	2	0	22		Yes	No
PGK1_72_2	10294	hfKM 02595	0	12	0	Yes		No
PGK1_72_2	10664	hfKM 05302	32	34	35	Yes	Yes	Possibly
PGK1_72_2	10764	hfKM 06044	24	27	23	Yes	Yes	Possibly
PGK1_72_2	10838	hfKM 06626	19	38	32	Yes	Yes	Possibly
PGK1_72_4	10838	hfKM 06626	1	8	3	Yes		No
PGK1_72_2	11044	hfKM 08103	9	15	10	Yes	Yes	No
PGK1_72_2	11133	hfKM 08617	35	36	38	Yes	Yes	Possibly
PGK1_72_2	12803	hfRF 01744	25	36	44	Yes	Yes	Possibly
PGK1_72_2	12920	hfRF 03771	78	17	4	Yes		Possibly
PGK1_72_2	13066	hfRH 00102	0	34	0	Yes		No
PGK1_72_2	13309	hfRH 01609	14	16	26	Yes	Yes	Possibly
PGK1_72_2	13670	hfRJC 02296	0	48	50	Yes	Yes	No
PGK1_72_2	13894	hfRJC 03897	2	16	0	Yes		No
PGK1_72_2	14488	hfS 13590	0	0	24		Yes	No
PGK1_72_2	14629	hfS 15380	42	50	40	Yes	Yes	Possibly
PGK1_72_2	14809	hfSCR 00731	12	30	16	Yes	Yes	Possibly
PGK1_72_2	15283	hfSEW 02660	0	0	26		Yes	No
PGK1_72_2	15394	hfSEW 03596	0	0	22		Yes	No
PGK1_72_2	16211	hfSPB 00468	2	22	32	Yes	Yes	No
PGK1_72_2	16212	hfSPB 00470	2	10	46	Yes	Yes	No
PGK1_72_2	16213	hfSPB 00471	30	31	32	Yes	Yes	Possibly
PGK1_72_2	16490	hfSPB 02620	25	27	31	Yes	Yes	Possibly
PGK1_72_2	17069	hfSPB 06981	2	21	16	Yes	Yes	No
PGK1_72_2	17170	hfSPB 07935	19	32	32	Yes	Yes	Possibly
PGK1_72_4	17327	hfXAX 00021	18	22	22	Yes	Yes	Possibly

PGK assay 2

	Eve ID	Maybridge ID	HsPGK	TbPGK	PvPGK	Tb active	Pv active	Toxic
PGK2_74_4	510		24	0	18		Yes	Possibly
PGK2_74_4	516		20	0	12		Yes	Possibly
PGK2_74_4	3259	hfBR 00086	14	0	8		Yes	Possibly
PGK2_74_2	3466	hfBTB 02152	0	4	8		Yes	No
PGK2_74_2	3594	hfBTB 02990	0	8	8	Yes	Yes	No
PGK2_74_2	3740	hfBTB 04150	0	2	8		Yes	No
PGK2_74_2	3951	hfBTB 05541	6	22	23	Yes	Yes	No
PGK2_74_2	4344	hfBTB 08470	0	0	10		Yes	No
PGK2_74_4	4584	hfBTB 10320	25	0	20		Yes	Possibly
PGK2_74_4	4585	hfBTB 10323	26	0	22		Yes	Possibly
PGK2_74_2	4655	hfBTB 11167	4	14	0	Yes		No
PGK2_74_2	6134	hfCD 11234	2	0	8		Yes	No
PGK2_74_4	6173	hfCD 11507	10	0	14		Yes	Possibly
PGK2_74_4	6365	hfDP 00986	10	0	10		Yes	Possibly
PGK2_74_2	7107	hfHTS 00290	0	8	12	Yes	Yes	No
PGK2_74_2	8028	hfHTS 05561	0	0	10		Yes	No
PGK2_74_2	8029	hfHTS 05567	0	0	8		Yes	No
PGK2_74_2	9499	hfJFD 00787	10	0	0			Possibly
PGK2_74_2	9504	hfJFD 00823	0	0	10		Yes	No
PGK2_74_4	10407	hfKM 03453	4	0	14		Yes	No
PGK2_74_2	10764	hfKM 06044	1	8	8	Yes	Yes	No
PGK2_74_2	12201	hfPD 00168	0	0	12		Yes	No
PGK2_74_2	12506	hfRDR 01446	0	10	20	Yes	Yes	No
PGK2_74_2	12803	hfRF 01744	12	16	13	Yes	Yes	Possibly
PGK2_74_4	12803	hfRF 01744	16	2	16		Yes	Possibly
PGK2_74_2	12830	hfRF 02175	1	0	8		Yes	No
PGK2_74_2	13309	hfRH 01609	0	8	4	Yes		No
PGK2_74_2	13670	hfRJC 02296	0	14	20	Yes	Yes	No
PGK2_74_2	14371	hfS 10607	13	18	15	Yes	Yes	Possibly
PGK2_74_2	14488	hfS 13590	0	0	12		Yes	No
PGK2_74_2	15661	hfSEW 04994	0	0	8		Yes	No
PGK2_74_2	16211	hfSPB 00468	0	0	8		Yes	No
PGK2_74_2	16212	hfSPB 00470	0	0	18		Yes	No
PGK2_74_4	17414		12	0	10		Yes	Possibly

A.9 Confirmation: active JHCCL compounds

DHFR TS3 assay

	Eve ID	JHCCL ID	HsDHFR	PvDHFR	PfRdhfr	Pv active	PfR active	Toxic
TS3_63_7	20110	SM_JHU-520	11	12	16	Yes	Yes	Possibly
TS3_63_7	20168	SM_JHU-904	8	8	22	Yes	Yes	Possibly
TS3_63_2	20168	SM_JHU-904	16	0	15		Yes	Possibly
TS3_63_3	20248	SM_JHU-1305	10	16	36	Yes	Yes	Possibly
TS3_63_7	20248	SM_JHU-1305	0	12	16	Yes	Yes	No
TS3_63_2	20414	SM_JHU-2095	8	32	1	Yes		No
TS3_63_7	20414	SM_JHU-2095	23	25	22	Yes	Yes	Possibly
TS3_63_2	20484	SM_JHU-2524	13	58	46	Yes	Yes	Possibly
TS3_63_2	21463	SM_JHU-9003	0	24	0	Yes		No

DHFR TS4 assay

	Eve ID	JHCCL ID	HsDHFR	TbDHFR	SmDHFR	Tb active	Sm active	Toxic
TS4_64_6	20248	SM_JHU-1305	8	8	24	Yes	Yes	Possibly
TS4_64_4	20516	SM_JHU-2766	0	4	12		Yes	No
TS4_64_4	20561	SM_JHU-3038	26	22	25	Yes	Yes	Possibly

DHFR TS5 assay

	Eve ID	JHCCL ID	PfRdhfr	HsDHFR	TcDHFR	PfR active	Tc active	Toxic
TS5_71_2	20110	SM_JHU-520	21	33	40	Yes	Yes	Possibly
TS5_71_5	20110	SM_JHU-520	11	17	18	Yes	Yes	Possibly
TS5_71_6	20110	SM_JHU-520	12	6	9		Yes	No
TS5_71_2	20168	SM_JHU-904	4	25	50		Yes	Possibly
TS5_71_5	20168	SM_JHU-904	7	22	26		Yes	Possibly
TS5_71_6	20168	SM_JHU-904	2	14	22		Yes	Possibly
TS5_71_3	20248	SM_JHU-1305	8	8	8	Yes	Yes	Possibly
TS5_71_5	20248	SM_JHU-1305	8	10	16	Yes	Yes	Possibly
TS5_71_6	20248	SM_JHU-1305	10	16	20	Yes	Yes	Possibly
TS5_71_6	20414	SM_JHU-2095	14	21	19	Yes	Yes	Possibly
TS5_71_2	20427	SM_JHU-2151	56	59	64	Yes	Yes	Possibly
TS5_71_2	20561	SM_JHU-3038	16	12	18	Yes	Yes	Possibly
TS5_71_5	21658	SM_JHU-10450	8	8	8	Yes	Yes	Possibly

DHFR TS6 assay

	Eve ID	JHCCL ID	HsDHFR	PvDHFR	PfDHFR	Pv active	Pf active	Toxic
TS6_77_11	20110	SM_JHU-520	17	11	14	Yes	Yes	Possibly
TS6_77_14	20110	SM_JHU-520	16	13	16	Yes	Yes	Possibly
TS6_c'pick	20141	SM_JHU-715	0	48	0	Yes		No
TS6_77_11	20168	SM_JHU-904	12	14	26	Yes	Yes	Possibly
TS6_77_16	20168	SM_JHU-904	14	13	18	Yes	Yes	Possibly
TS6_77_14	20168	SM_JHU-904	4	3	18		Yes	No
TS6_c'pick	20245	SM_JHU-1293	0	40	0	Yes		No
TS6_c'pick	20248	SM_JHU-1305	12	44	50	Yes	Yes	Possibly
TS6_77_11	20248	SM_JHU-1305	8	16	16	Yes	Yes	Possibly
TS6_77_14	20248	SM_JHU-1305	0	12	12	Yes	Yes	No
TS6_c'pick	20414	SM_JHU-2095	6	30	8	Yes		No
TS6_77_14	20414	SM_JHU-2095	20	24	24	Yes	Yes	Possibly
TS6_77_16	20414	SM_JHU-2095	22	24	21	Yes	Yes	Possibly
TS6_c'pick	20484	SM_JHU-2524	16	48	46	Yes	Yes	Possibly
TS6_c'pick	20516	SM_JHU-2766	10	20	16	Yes	Yes	Possibly
TS6_c'pick	20940	SM_JHU-5529	0	44	0	Yes		No
TS6_c'pick	21389	SM_JHU-8509	0	80	80	Yes	Yes	No
TS6_c'pick	21438	SM_JHU-8859	0	10	12	Yes	Yes	No
TS6_c'pick	21463	SM_JHU-9003	0	12	0	Yes		No
TS6_c'pick	21625	SM_JHU-10251	0	38	0	Yes		No
TS6_c'pick	21658	SM_JHU-10450	20	40	23	Yes	Yes	Possibly
TS6_77_14	21658	SM_JHU-10450	8	10	8	Yes	Yes	Possibly
TS6_77_16	21658	SM_JHU-10450	8	8	6	Yes		Possibly
TS6_77_11	21658	SM_JHU-10450	8	7	8		Yes	Possibly
TS6_77_16	21767	SM_JHU-12096	7	11	11	Yes	Yes	No
TS6_c'pick	21767	SM_JHU-12096	8	9	12		Yes	No
TS6_77_14	21919		0	10	0	Yes		No

DHFR TS7 assay

	Eve ID	JHCCL ID	HsDHFR	LmDHFR	PvRdhfr	Lm active	PvR active	Toxic
TS7_80_4	20110	SM_JHU-520	8	8	0	Yes		Possibly
TS7_80_6	20110	SM_JHU-520	22	20	0	Yes		Possibly
TS7_80_3	20134	SM_JHU-657	20	12	0	Yes		Possibly
TS7_80_4	20248	SM_JHU-1305	8	8	12	Yes	Yes	Possibly
TS7_80_6	20248	SM_JHU-1305	0	8	10	Yes	Yes	No
TS7_80_3	20449	SM_JHU-2317	40	40	0	Yes		Possibly
TS7_80_3	20472	SM_JHU-2438	22	16	0	Yes		Possibly
TS7_80_3	21619	SM_JHU-10190	40	40	16	Yes	Yes	Possibly
TS7_80_6	21658	SM_JHU-10450	8	8	4	Yes		Possibly
TS7_80_3	21767	SM_JHU-12096	0	0	16		Yes	No
TS7_80_6	21767	SM_JHU-12096	32	36	32	Yes	Yes	Possibly

NMT assay 1

	Eve ID	JHCCL ID	HsNMT	TbNMT	PvNMT	Tb active	Pv active	Toxic
NMT_78_6	20110	SM_JHU-520	14	8	8	Yes	Yes	Possibly
NMT_78_4	20248	SM_JHU-1305	16	24	24	Yes	Yes	Possibly
NMT_78_6	20248	SM_JHU-1305	4	18	12	Yes	Yes	No
NMT_78_6	20414	SM_JHU-2095	16	19	19	Yes	Yes	Possibly
NMT_78_2	20470	SM_JHU-2430	1	24	20	Yes	Yes	No
NMT_78_2	20484	SM_JHU-2524	4	28	26	Yes	Yes	No
NMT_78_2	20548	SM_JHU-2987	8	16	16	Yes	Yes	Possibly
NMT_78_2	20626	SM_JHU-3353	0	32	2	Yes		No
NMT_78_2	20633	SM_JHU-3415	0	10	0	Yes		No
NMT_78_2	21006	SM_JHU-6047	0	32	0	Yes		No
NMT_78_6	21006	SM_JHU-6047	0	40	30	Yes	Yes	No
NMT_78_2	21400	SM_JHU-8555	1	20	14	Yes	Yes	No
NMT_78_6	21658	SM_JHU-10450	6	8	7	Yes		No
NMT_78_6	21919		0	22	0	Yes		No

NMT assay 2

	Eve ID	JHCCL ID	HsNMT	SmNMT	TcNMT	Sm active	Tc active	Toxic
NMT_79_6	20110	SM_JHU-520	16	11	11	Yes	Yes	Possibly
NMT_79_6	20168	SM_JHU-904	3	0	17		Yes	No
NMT_79_2	20245	SM_JHU-1293	0	8	10	Yes	Yes	No
NMT_79_6	20248	SM_JHU-1305	8	14	12	Yes	Yes	Possibly
NMT_79_6	20414	SM_JHU-2095	20	21	20	Yes	Yes	Possibly
NMT_79_2	20449	SM_JHU-2317	40	40	40	Yes	Yes	Possibly
NMT_79_2	20472	SM_JHU-2438	26	24	20	Yes	Yes	Possibly
NMT_79_2	20525	SM_JHU-2831	0	16	0	Yes		No
NMT_79_2	20626	SM_JHU-3353	4	40	40	Yes	Yes	No
NMT_79_6	21006	SM_JHU-6047	0	40	0	Yes		No
NMT_79_2	21006	SM_JHU-6047	0	40	40	Yes	Yes	No
NMT_79_2	21281	SM_JHU-7761	0	0	16		Yes	No
NMT_79_2	21390	SM_JHU-8513	10	8	10	Yes	Yes	Possibly
NMT_79_2	21463	SM_JHU-9003	0	2	32		Yes	No
NMT_79_6	21658	SM_JHU-10450	8	8	8	Yes	Yes	Possibly
NMT_79_6	21919		12	16	10	Yes	Yes	Possibly

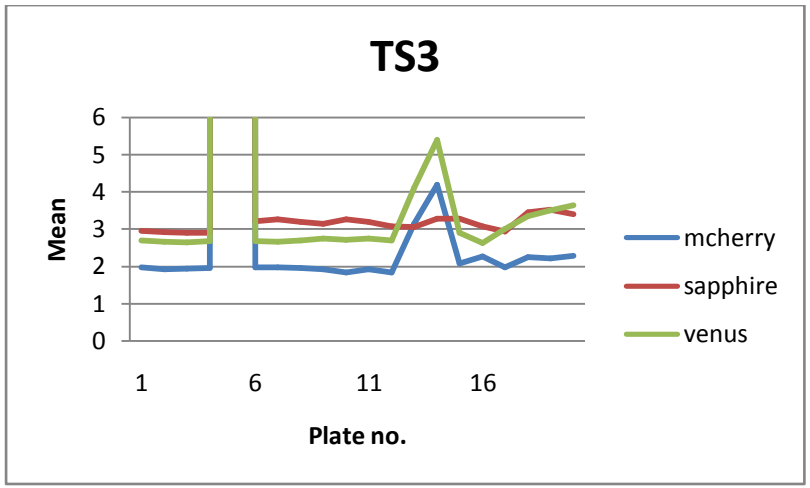
PGK assay 1

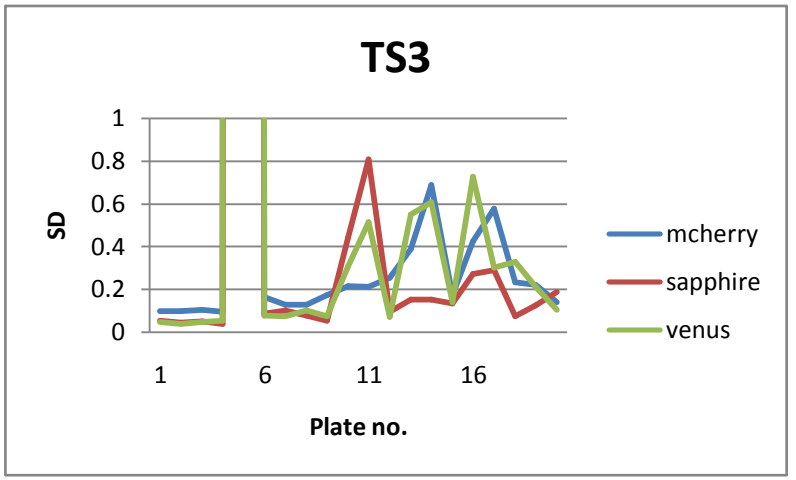
	Eve ID	JHCCL ID	HsPGK	SmPGK	TcPGK	Sm active	Tc active	Toxic
PGK1_72_2	20070	SM_JHU-327	30	0	26		Yes	Possibly
PGK1_72_2	20110	SM_JHU-520	31	32	30	Yes	Yes	Possibly
PGK1_72_2	20460	SM_JHU-2381	30	29	29	Yes	Yes	Possibly
PGK1_72_2	20635	SM_JHU-3437	78	80	25	Yes	Yes	Possibly
PGK1_72_2	20655	SM_JHU-3636	6	0	20		Yes	No
PGK1_72_2	20750	SM_JHU-4214	64	74	72	Yes	Yes	Possibly
PGK1_72_2	21088	SM_JHU-6452	80	80	18	Yes	Yes	Possibly
PGK1_72_2	21178	SM_JHU-7015	28	64	25	Yes	Yes	Possibly
PGK1_72_2	21224	SM_JHU-7333	70	80	32	Yes	Yes	Possibly

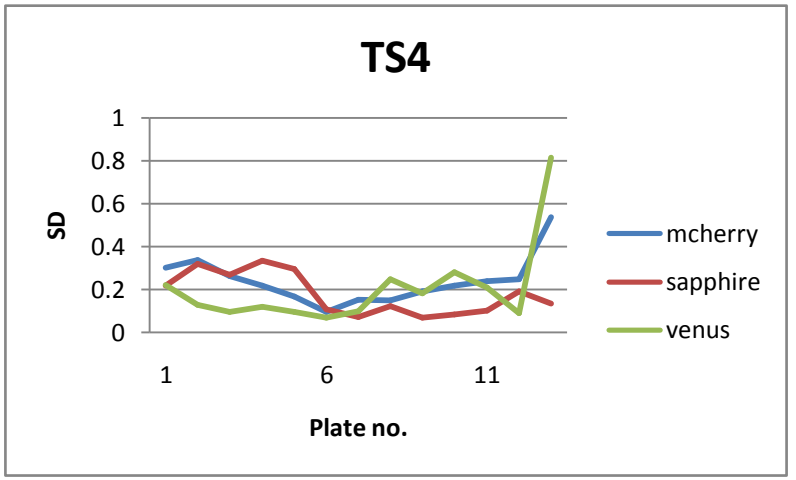
PGK assay 2

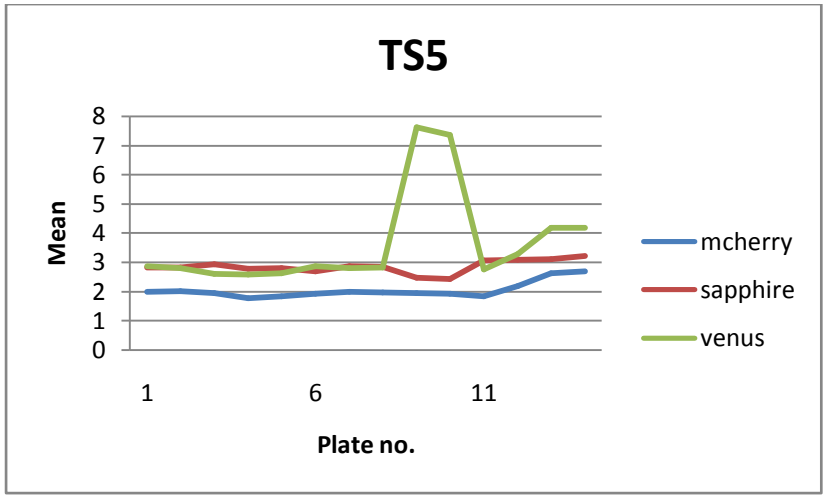
	Eve ID	JHCCL ID	HsPGK	TbPGK	PvPGK	Tb active	Pv active	Toxic
PGK2_74_2	20110	SM_JHU-520	6	8	7	Yes		No
PGK2_74_2	20134	SM_JHU-657	17	16	16	Yes	Yes	Possibly
PGK2_74_2	20291	SM_JHU-1539	40	40	40	Yes	Yes	Possibly
PGK2_74_2	20427	SM_JHU-2151	26	24	24	Yes	Yes	Possibly
PGK2_74_2	20750	SM_JHU-4214	37	33	32	Yes	Yes	Possibly
PGK2_74_2	21088	SM_JHU-6452	40	38	22	Yes	Yes	Possibly
PGK2_74_2	21224	SM_JHU-7333	40	40	38	Yes	Yes	Possibly

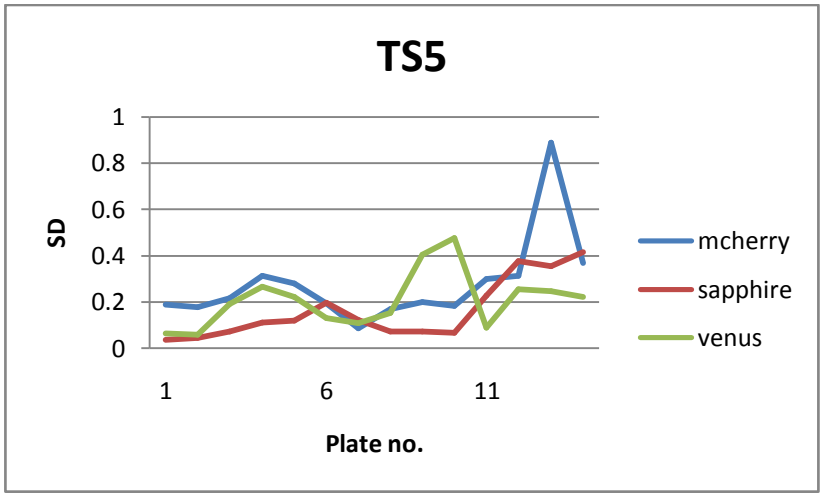
A.10 Confirmation screens: negative control statistics for doubling time

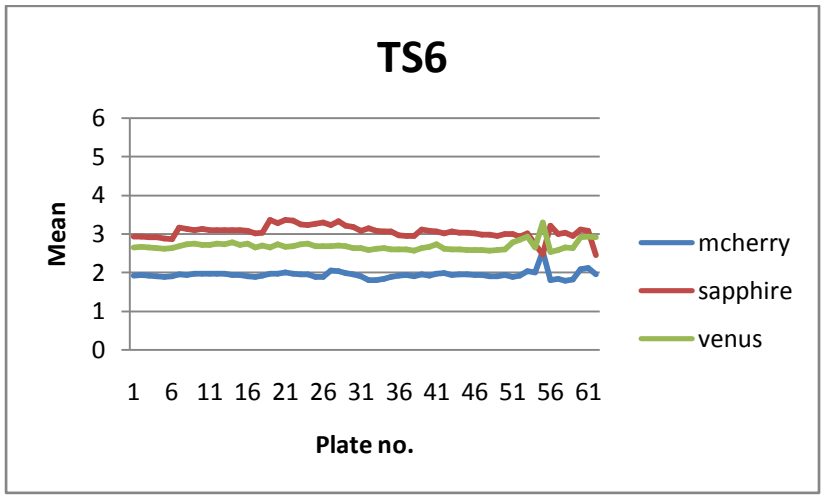


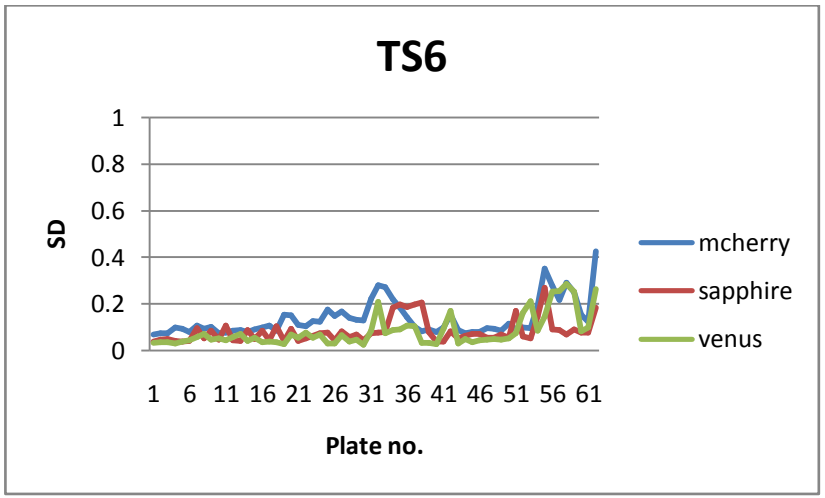


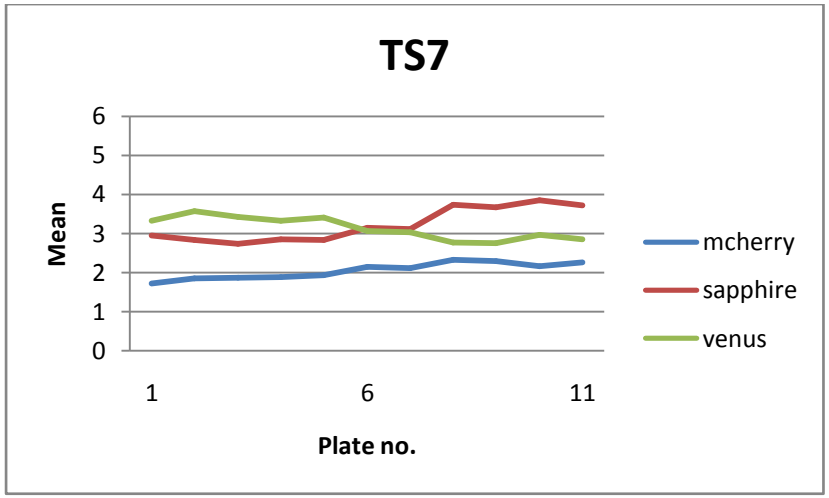


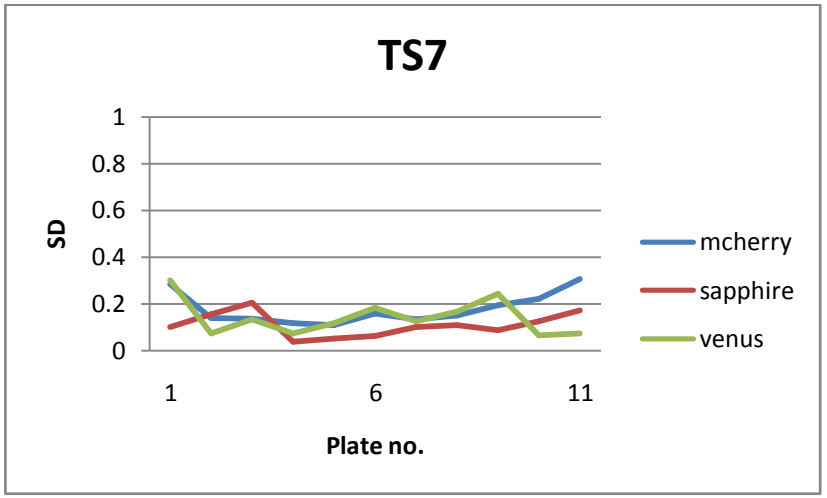

TS5
Mean
8
7
6
5
4
3
2
1
0
1
6
11
Plate no.
mcherry
sapphire
venus

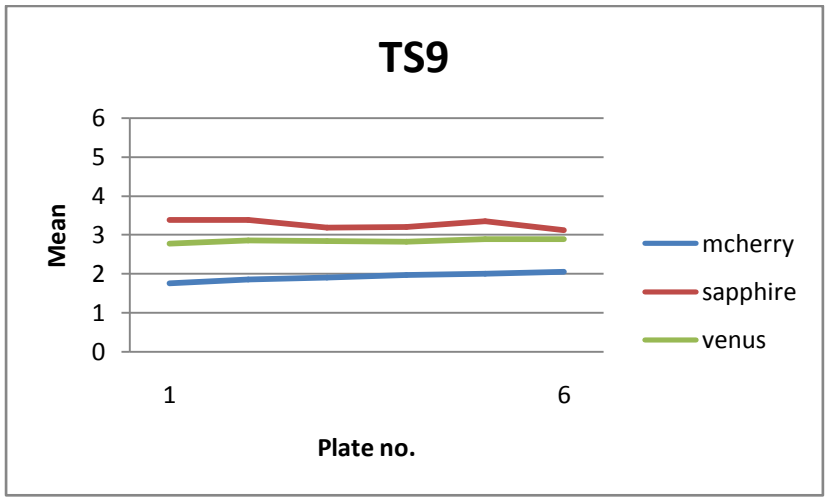

TS5
SD
1
0.8
0.6
0.4
0.2
0
1
6
11
Plate no.
mcherry
sapphire
venus

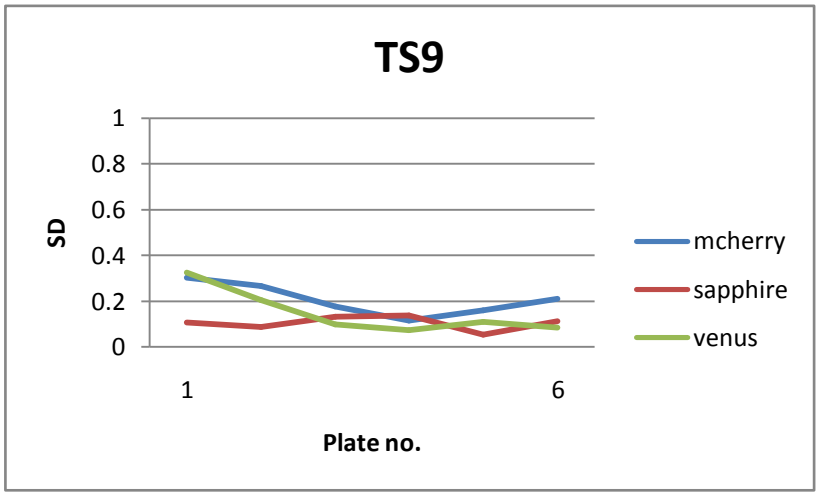

TS6
Mean
6
5
4
3
2
1
0
1 6 11 16 21 26 31 36 41 46 51 56 61
Plate no.
mcherry
sapphire
venus

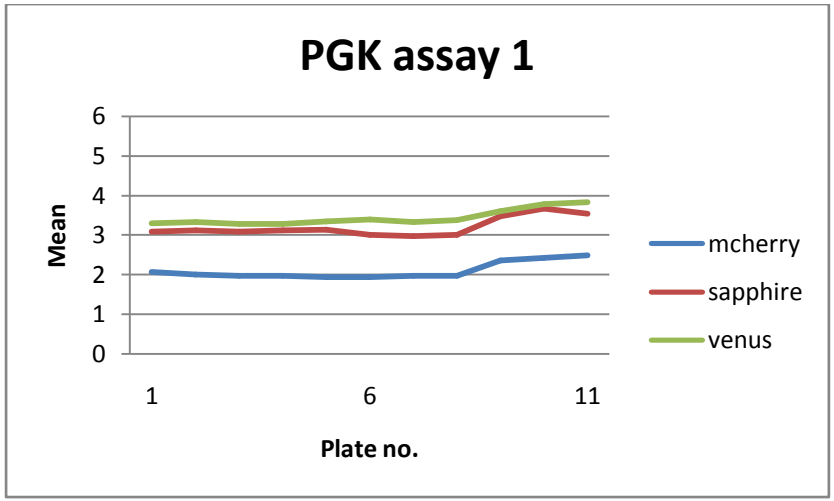

TS6
SD
1
0.8
0.6
0.4
0.2
0
1 6 11 16 21 26 31 36 41 46 51 56 61
Plate no.
mcherry
sapphire
venus

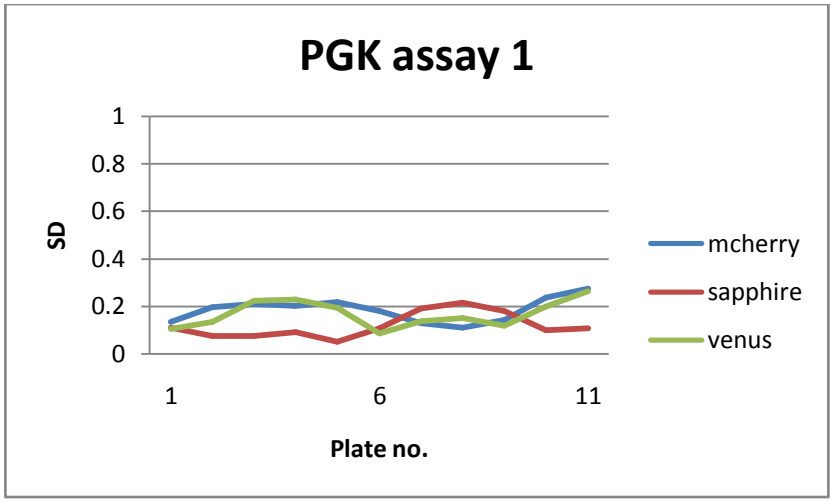

TS7
Mean
6
5
4
3
2
1
0
1
6
11
Plate no.
mcherry
sapphire
venus

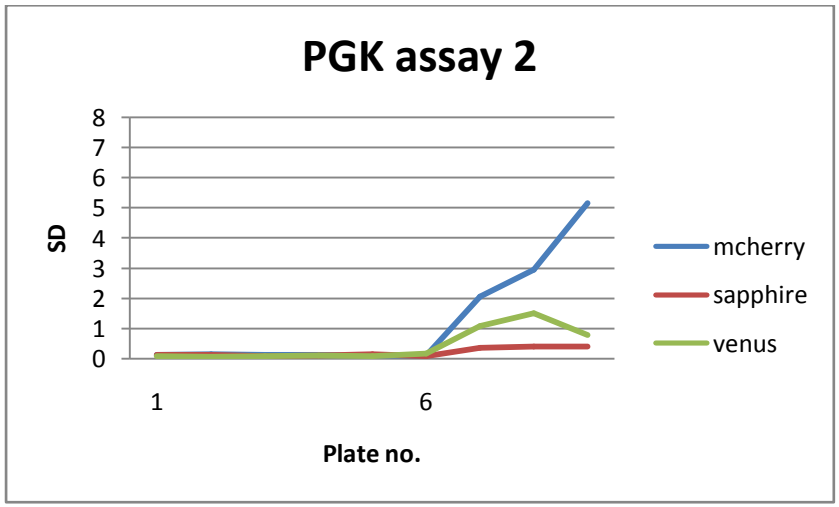

TS7
SD
1
0.8
0.6
0.4
0.2
0
1
6
11
Plate no.
mcherry
sapphire
venus

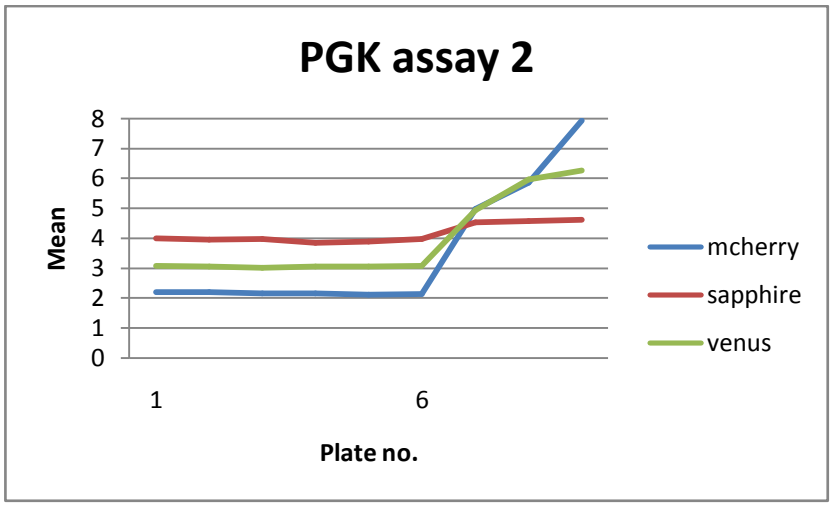

TS9
Mean
6
5
4
3
2
1
0
1
6
Plate no.
mcherry
sapphire
venus

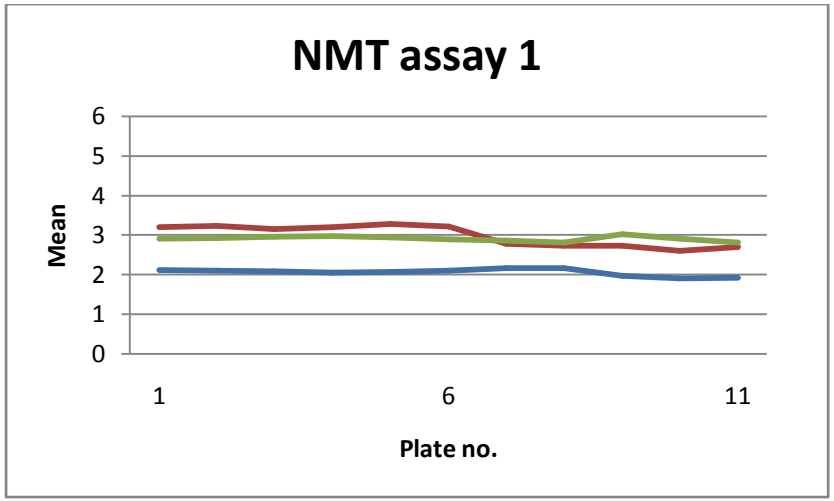

TS9
SD
1
0.8
0.6
0.4
0.2
0
1
6
Plate no.
mcherry
sapphire
venus

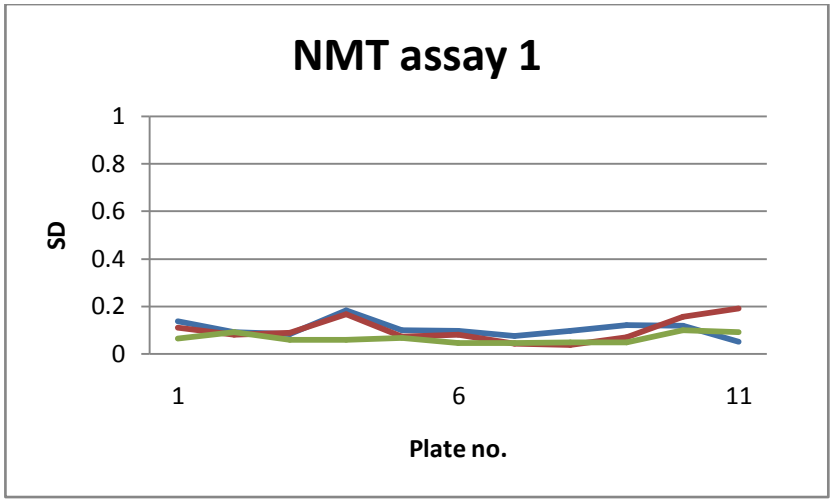

PGK assay 1
Mean
6
5
4
3
2
1
0
1
6
11
Plate no.
mcherry
sapphire
venus

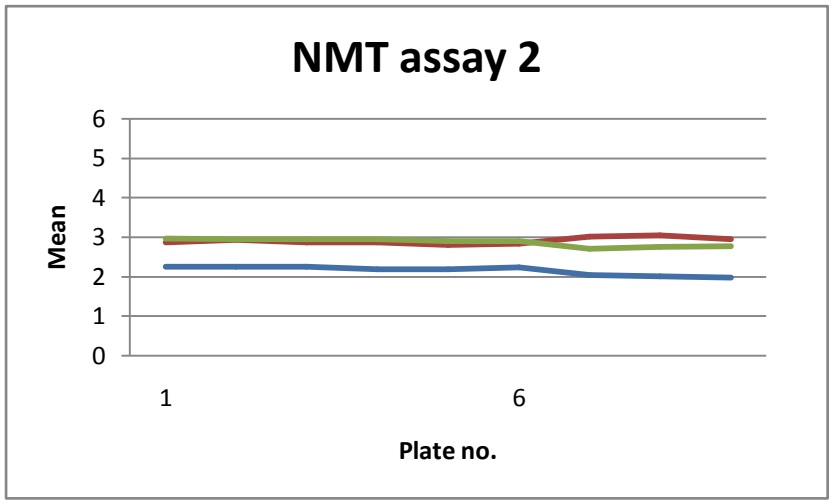

PGK assay 1
SD
1
0.8
0.6
0.4
0.2
0
1
6
11
Plate no.
mcherry
sapphire
venus

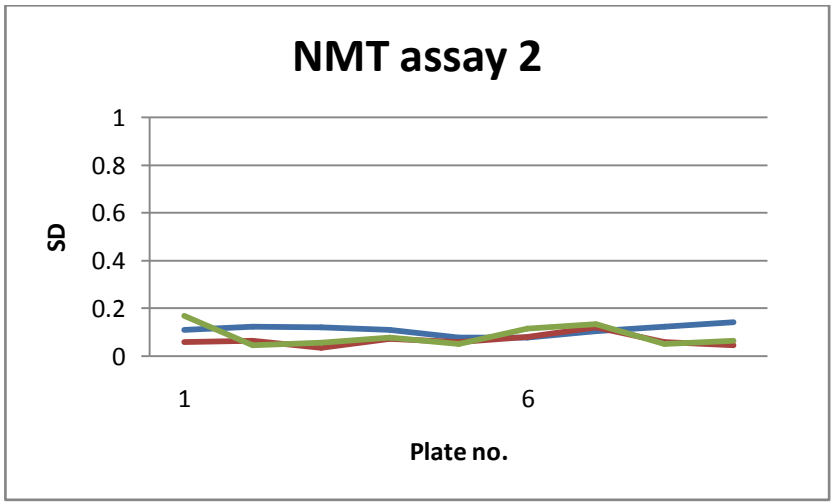

PGK assay 2
SD
8
7
6
5
4
3
2
1
0
1
6
Plate no.
mcherry
sapphire
venus


PGK assay 2
Mean
8
7
6
5
4
3
2
1
0
1
6
Plate no.
mcherry
sapphire
venus


NMT assay 1
Mean
6
5
4
3
2
1
0
1
6
11
Plate no.


NMT assay 1
SD
1
0.8
0.6
0.4
0.2
0
1
6
11
Plate no.


NMT assay 2
Mean
6
5
4
3
2
1
0
1
6
Plate no.


NMT assay 2
SD
1
0.8
0.6
0.4
0.2
0
1
6
Plate no.

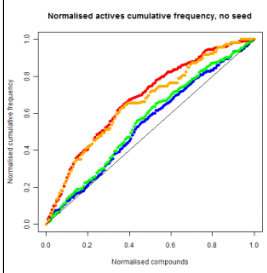
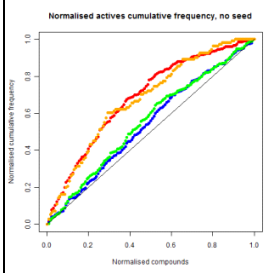
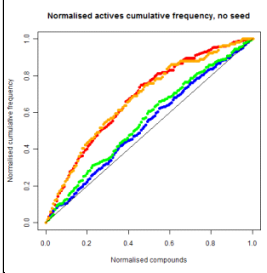
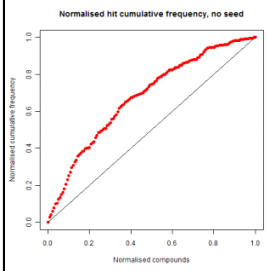
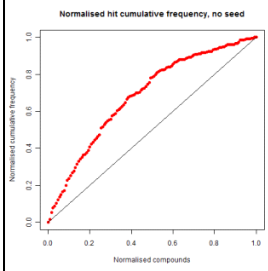
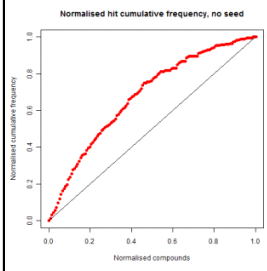
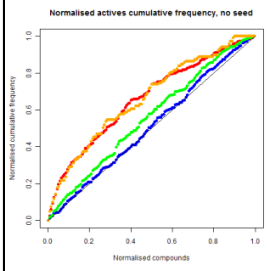
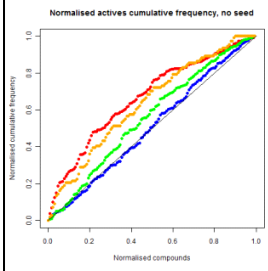
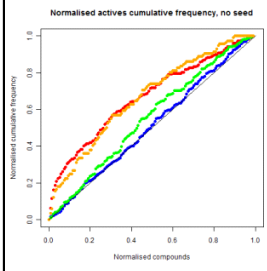
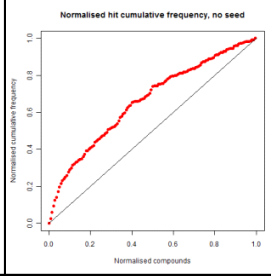
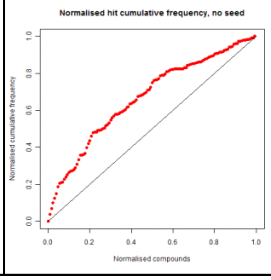
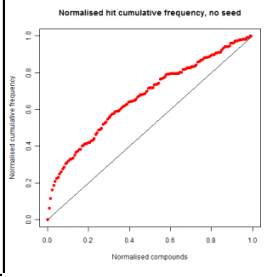
Appendix B

Simulated Active Learning curves and rare category detection

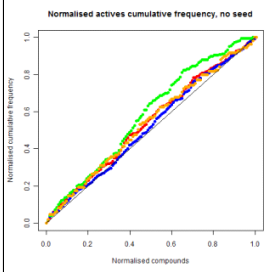
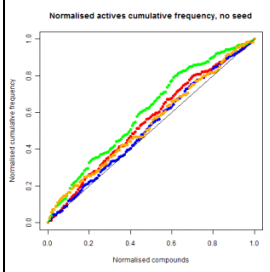
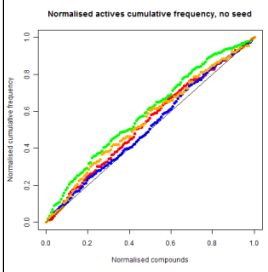
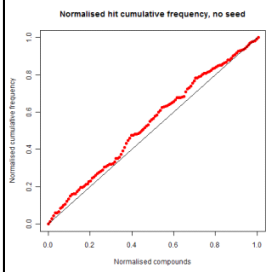
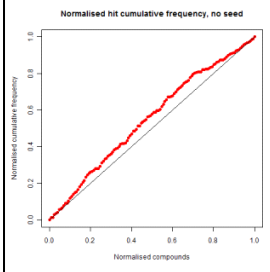
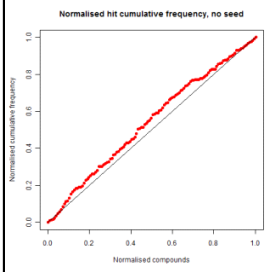
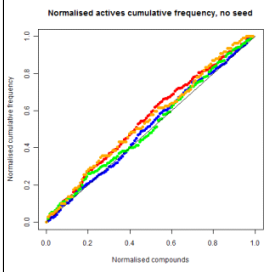
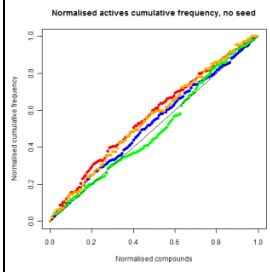
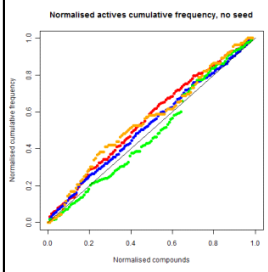
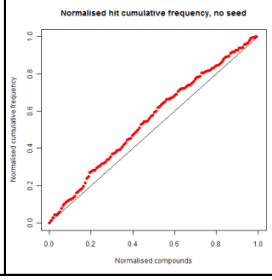
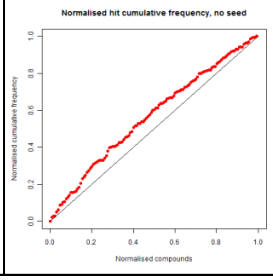
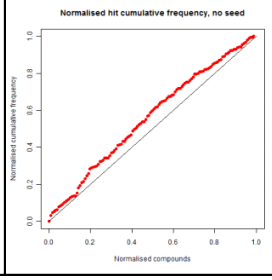
Descriptions and keys for the Active Learning curves and deficiency measurements are given in Section 5.1 of the main body of the thesis.

B.1 Active k-optimisation and SimplyGreedy learning curves

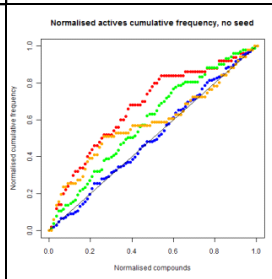
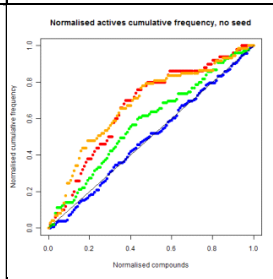
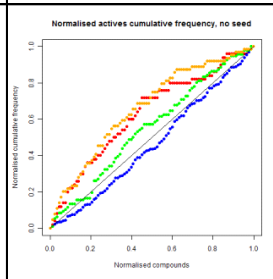
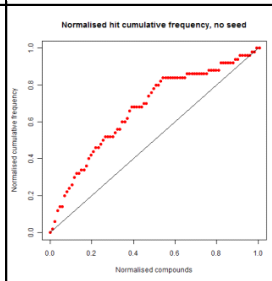
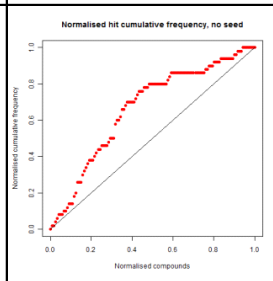
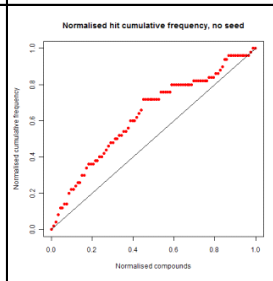
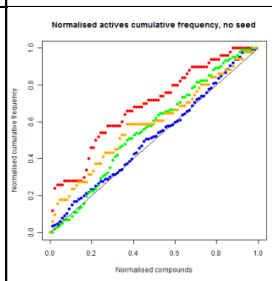
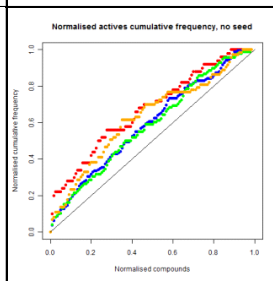
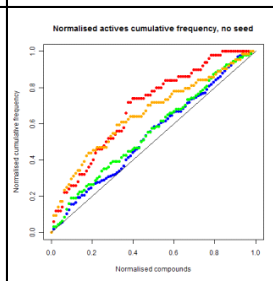
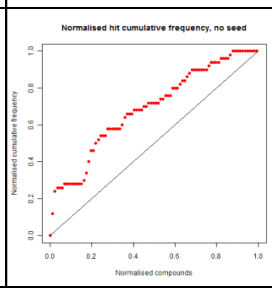
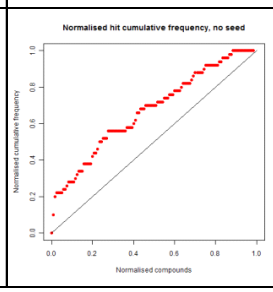
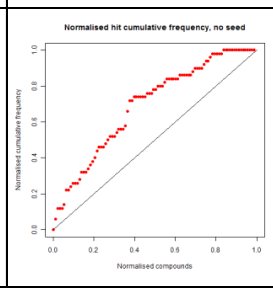
TS3 PvDHFR

Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

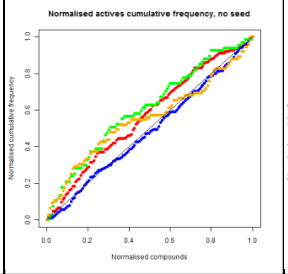
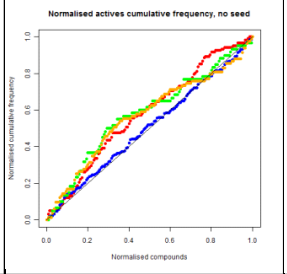
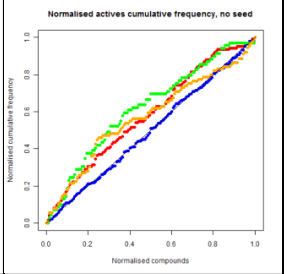
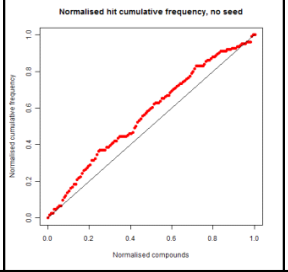
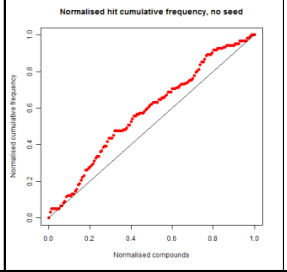
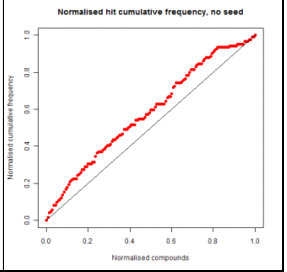
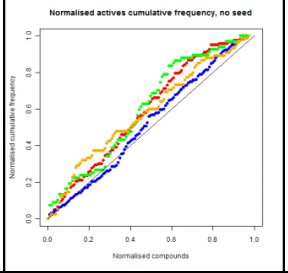
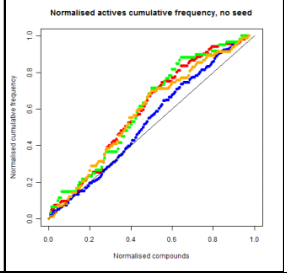
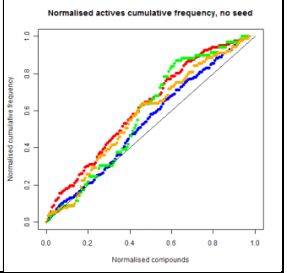
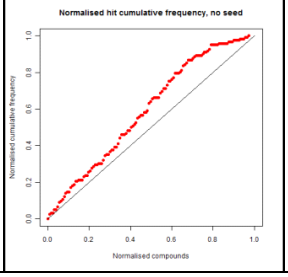
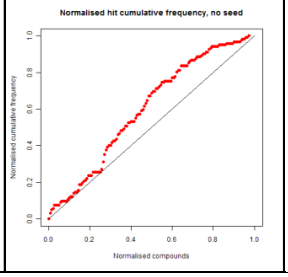
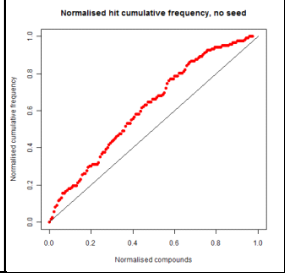
TS3 PfRdhfr

Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

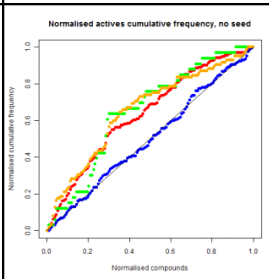
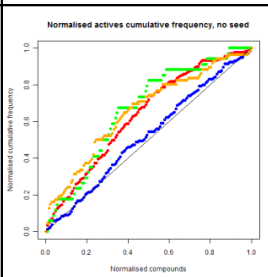
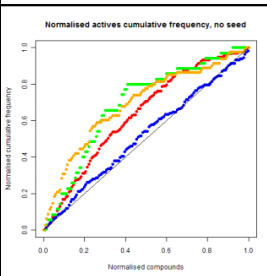
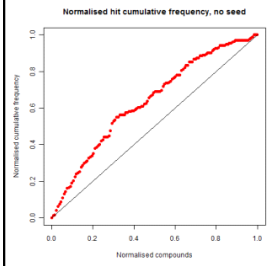
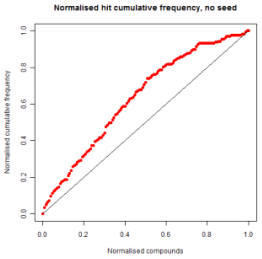
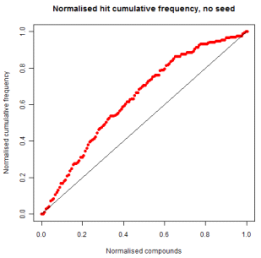
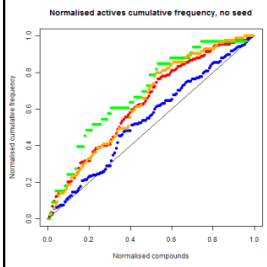
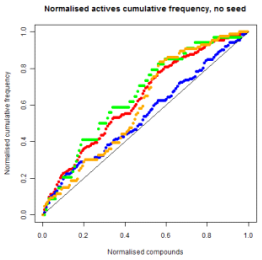
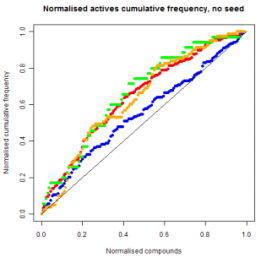
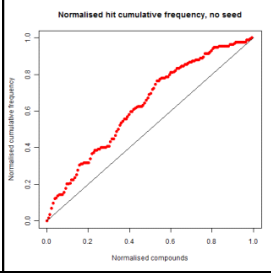
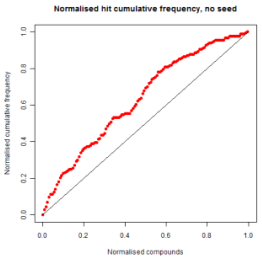
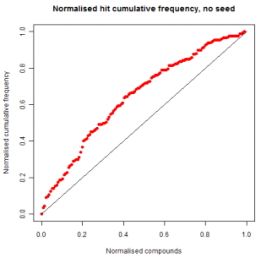
TS4 TbDHFR

Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

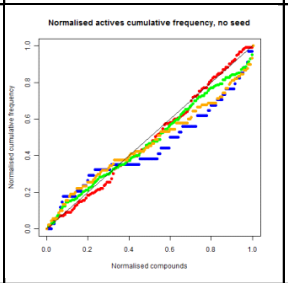
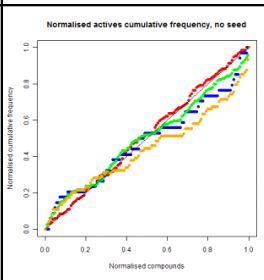
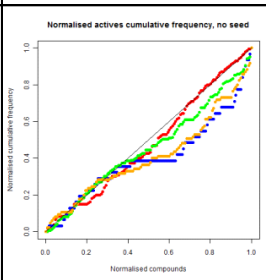
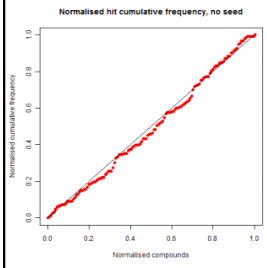
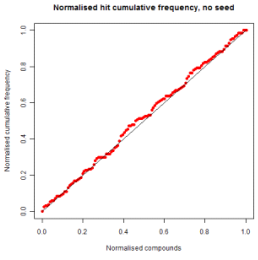
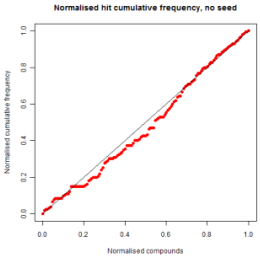
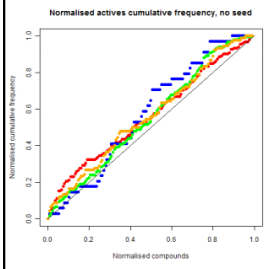
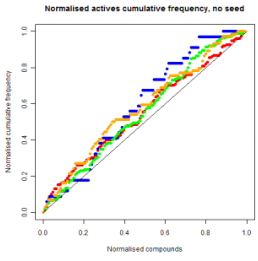
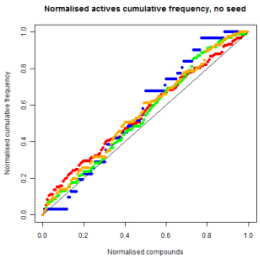
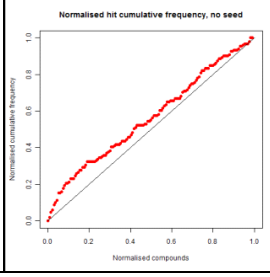
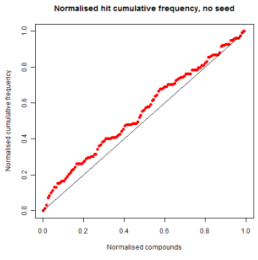
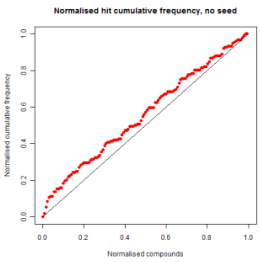
TS4 SmDHFR

Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

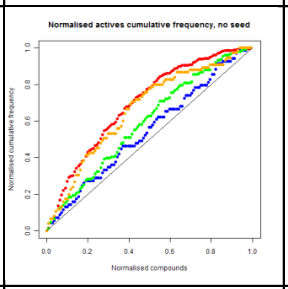
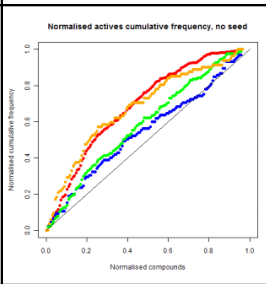
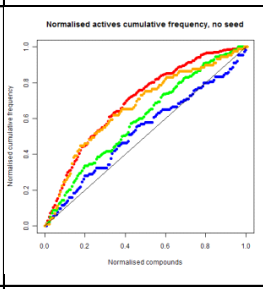
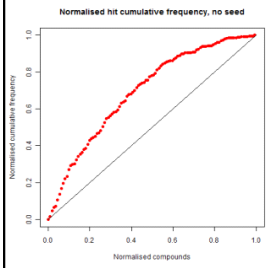
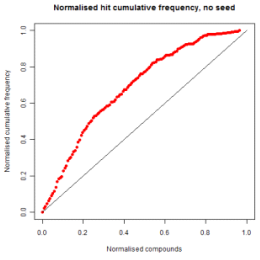
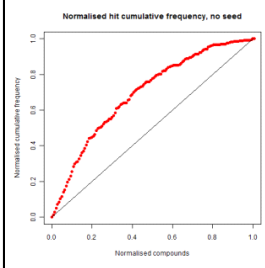
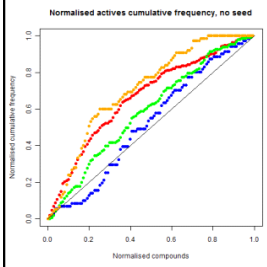
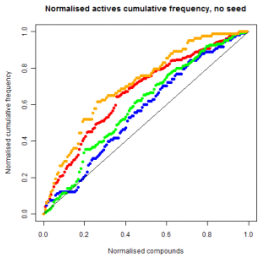
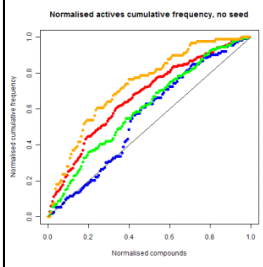
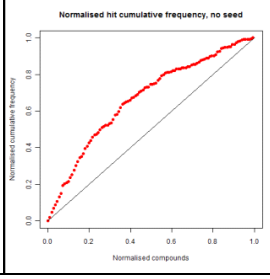
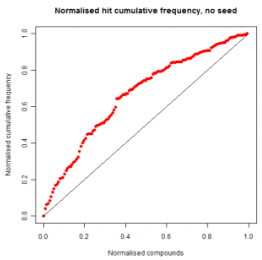
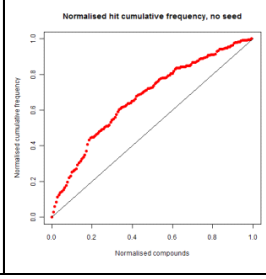
TS5 TcDHFR

Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

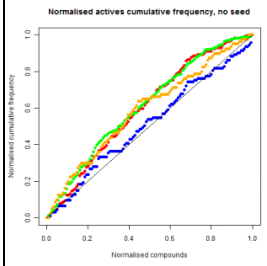
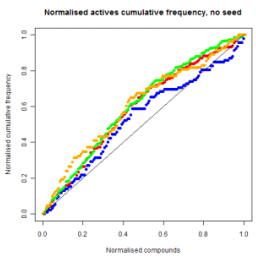
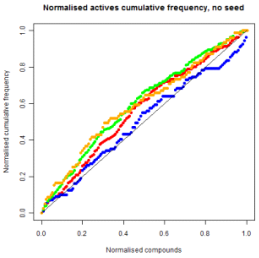
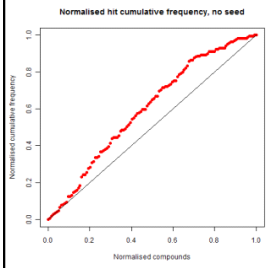
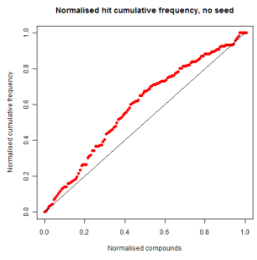
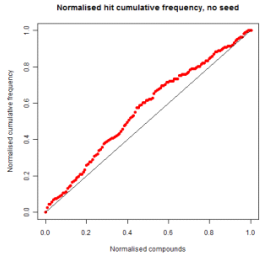
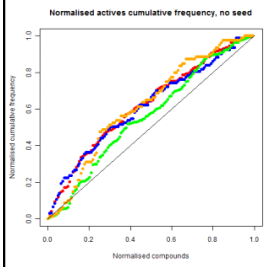
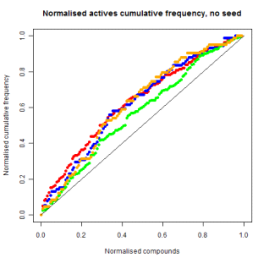
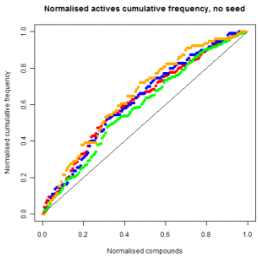
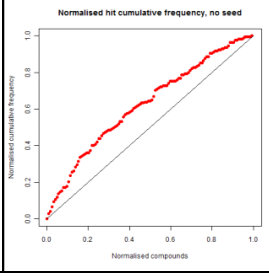
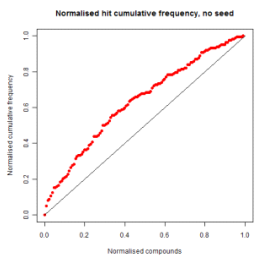
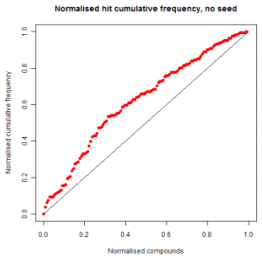
TS5 PfRdhfr

Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

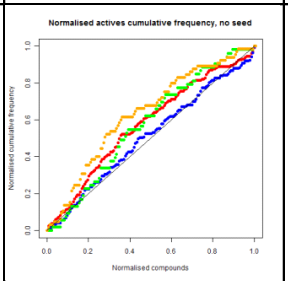
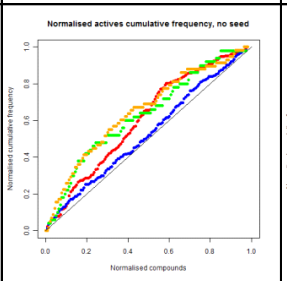
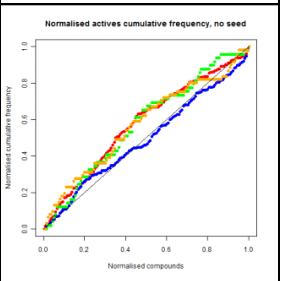
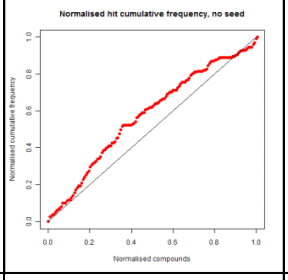
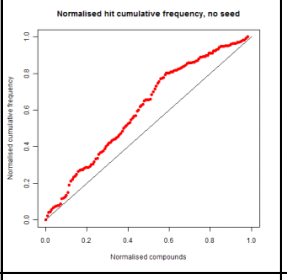
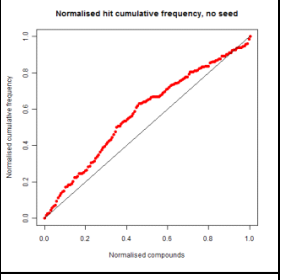
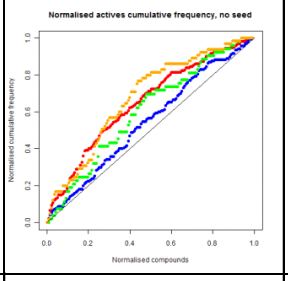
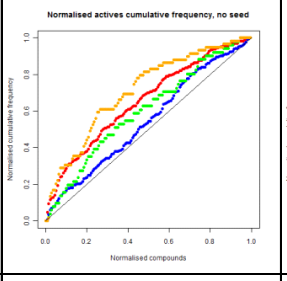
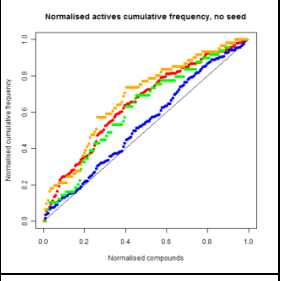
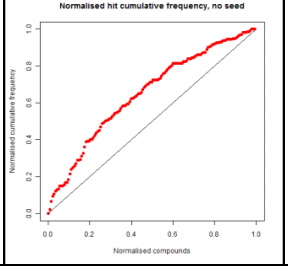
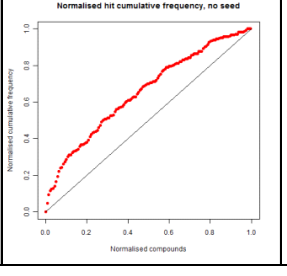
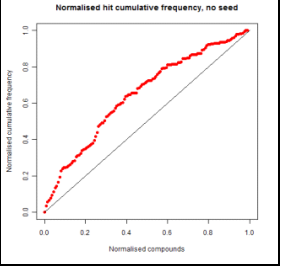
TS6 PvDHFR (there were 8 other runs for each algorithm)

Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

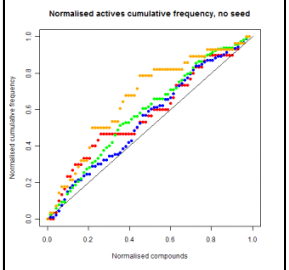
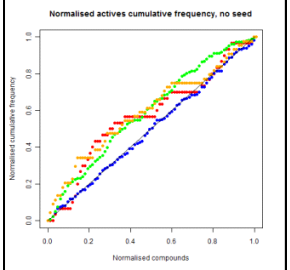
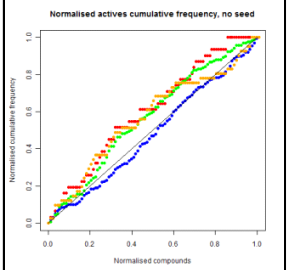
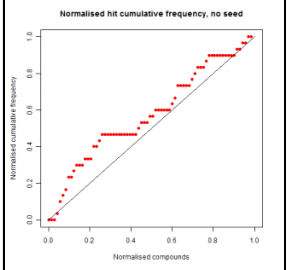
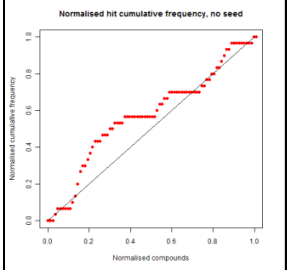
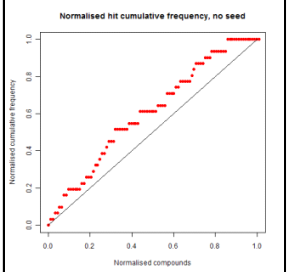
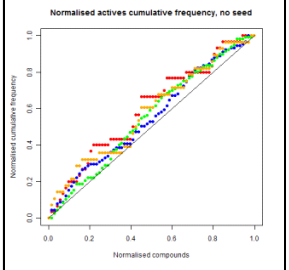
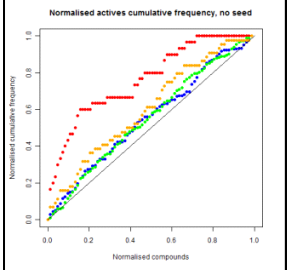
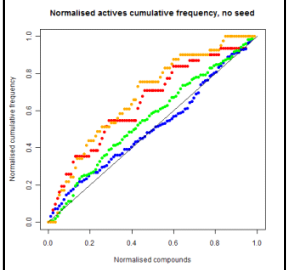
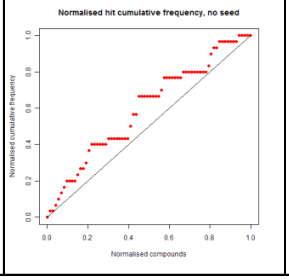
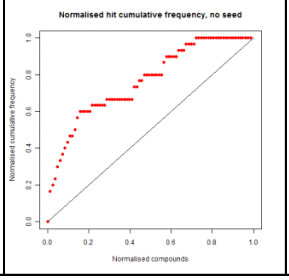
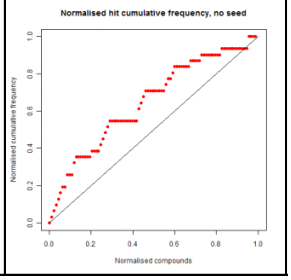
TS6 PfDHFR

Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

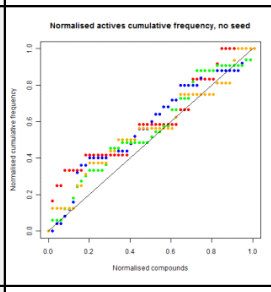
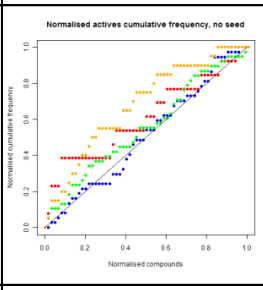
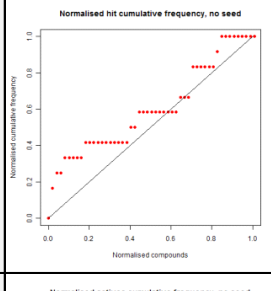
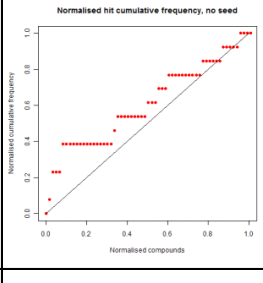
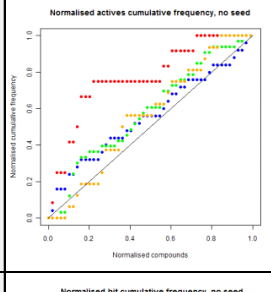
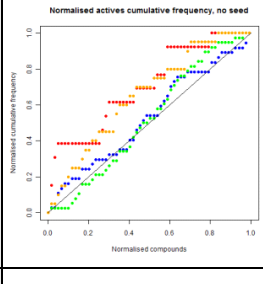
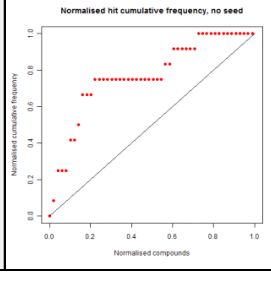
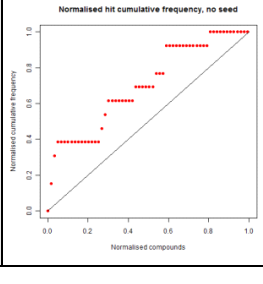
TS7 PvRdhfr

Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

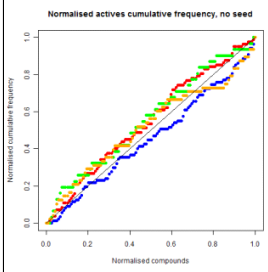
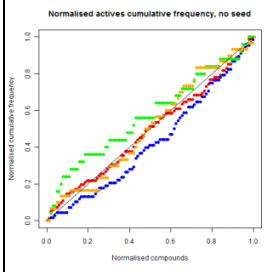
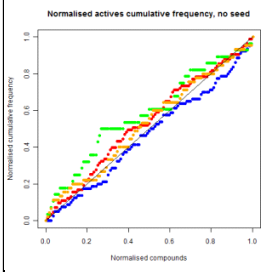
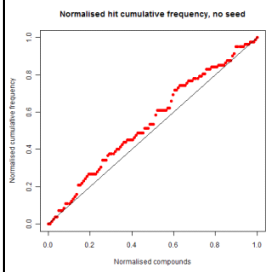
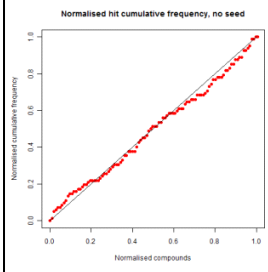
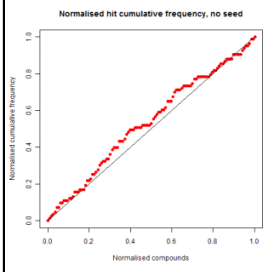
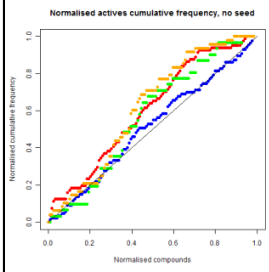
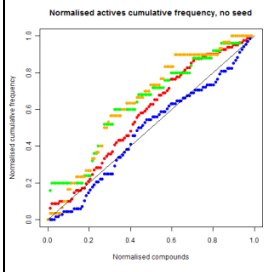
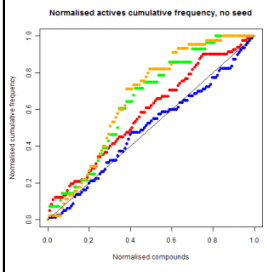
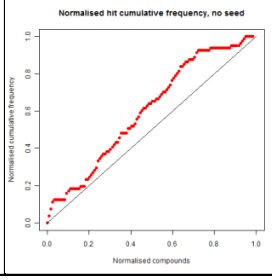
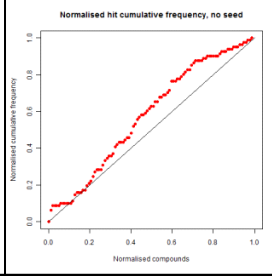
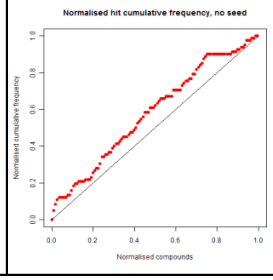
TS7 LmDHFR

Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

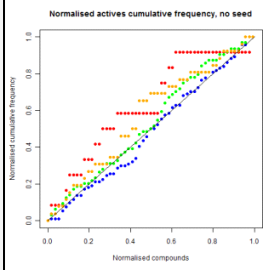
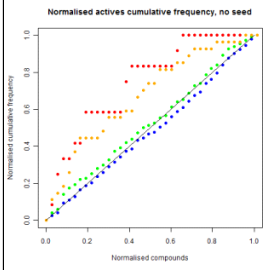
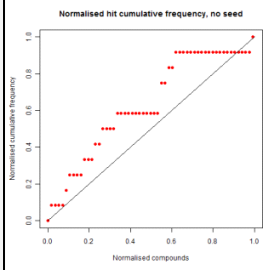
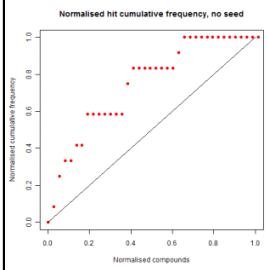
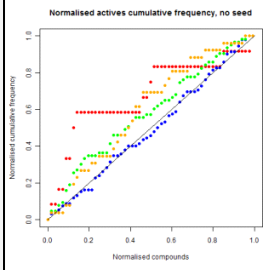
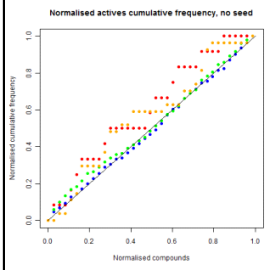
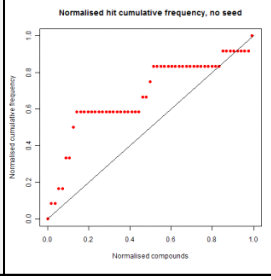
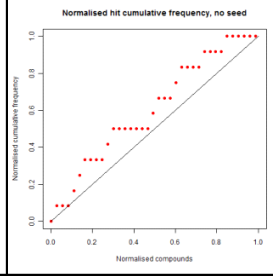
PGK-1 TbPGK

Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

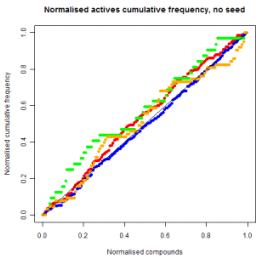
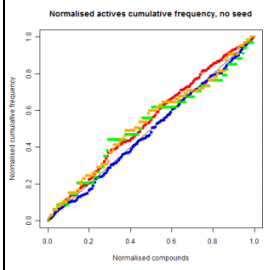
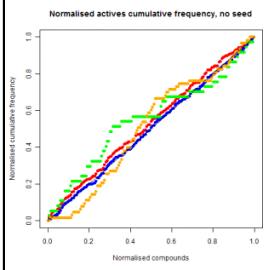
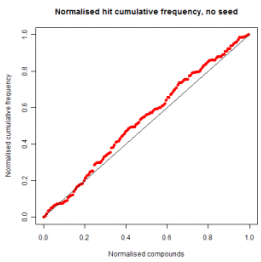
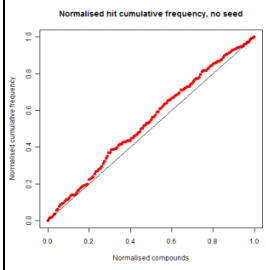
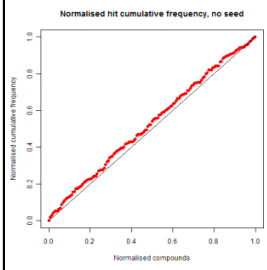
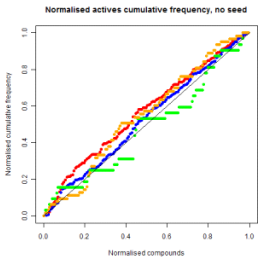
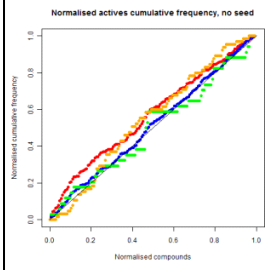
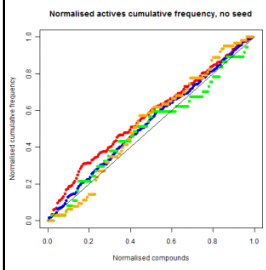
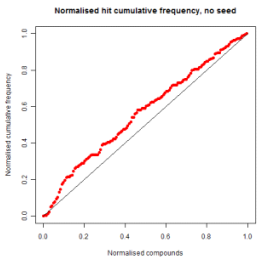
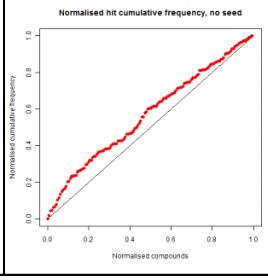
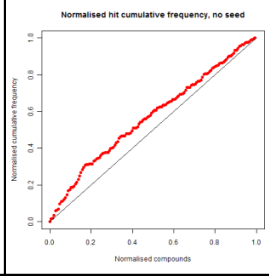
PGK-1 PvPGK

Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

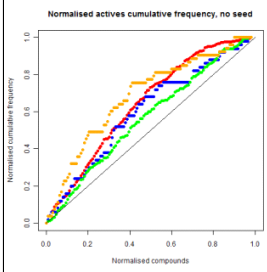
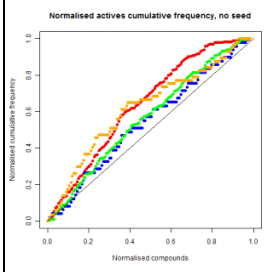
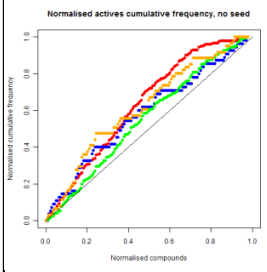
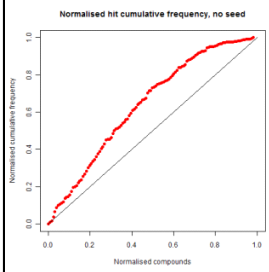
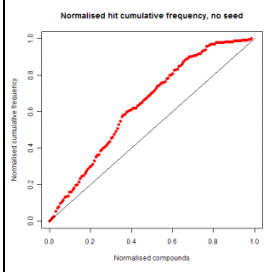
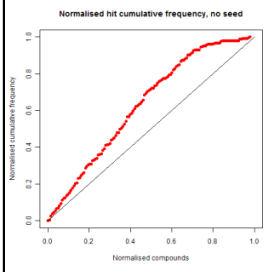
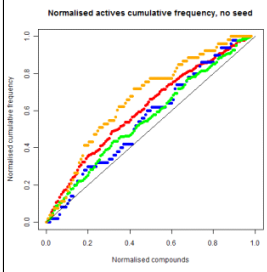
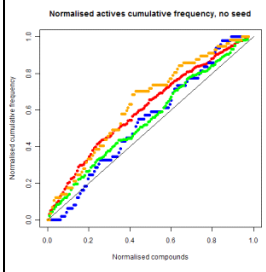
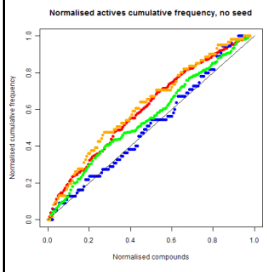
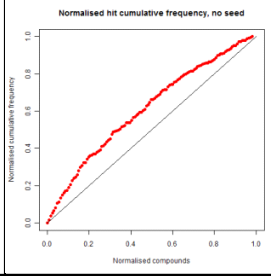
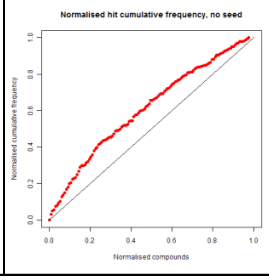
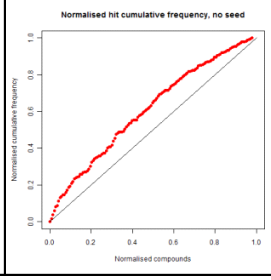
PGK-2 SmPGK

Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

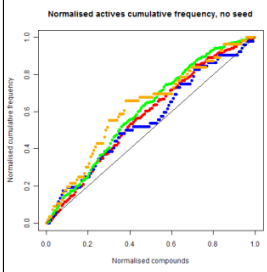
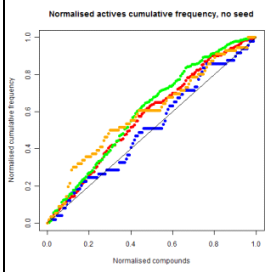
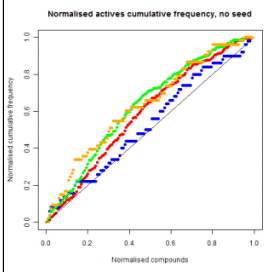
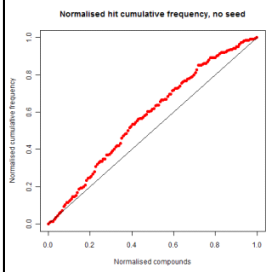
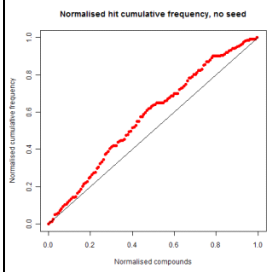
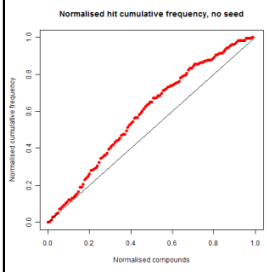
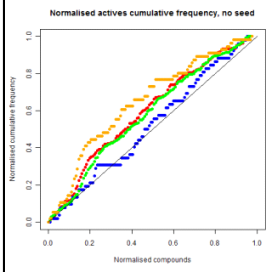
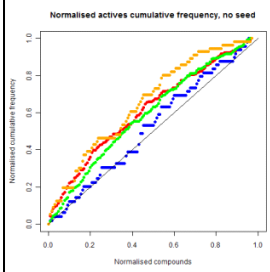
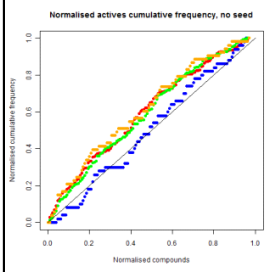
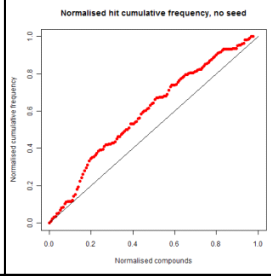
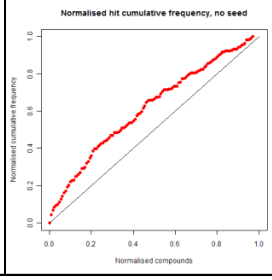
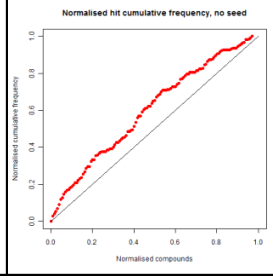
PGK-2 TcPGK

Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

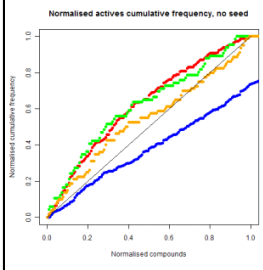
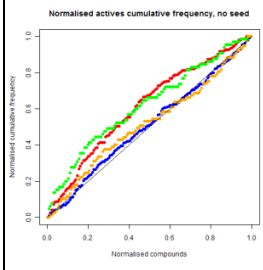
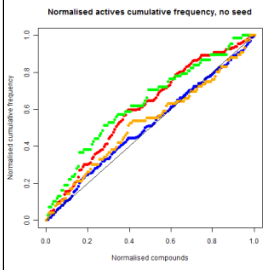
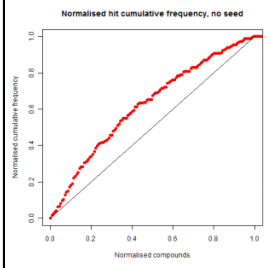
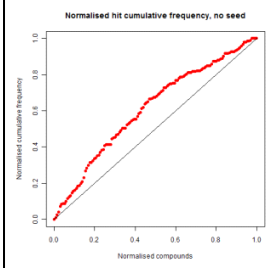
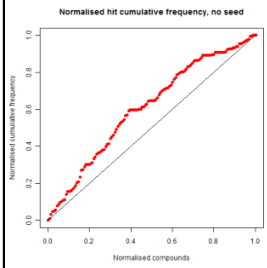
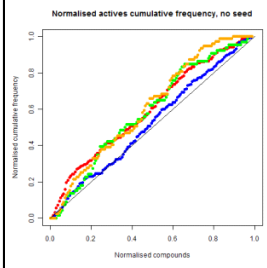
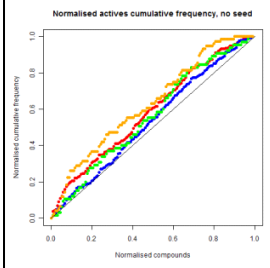
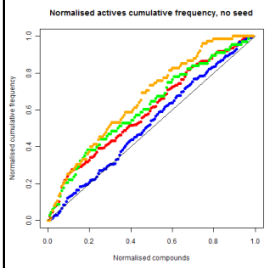
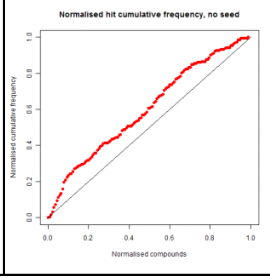
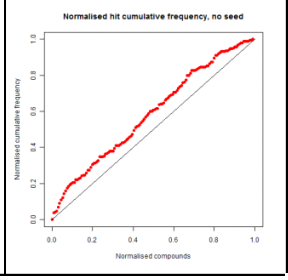
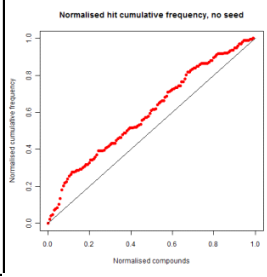
NMT-1 TbNMT

Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

NMT-1 PvNMT

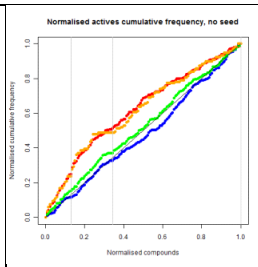
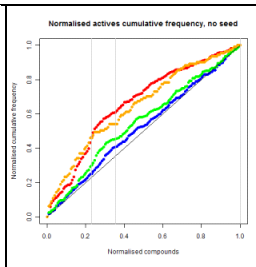
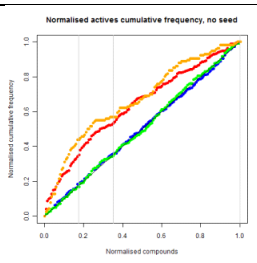
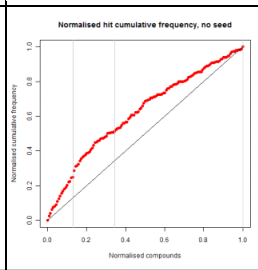
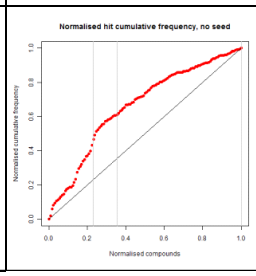
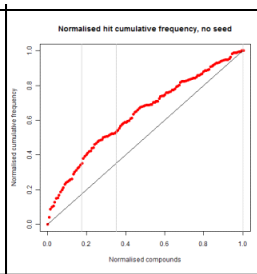
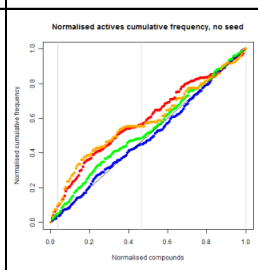
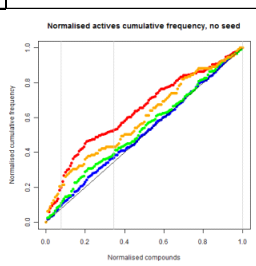
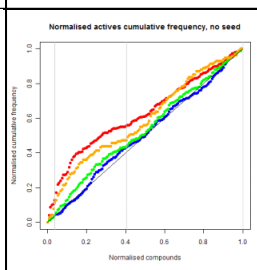
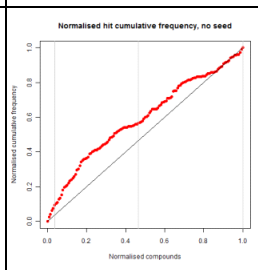
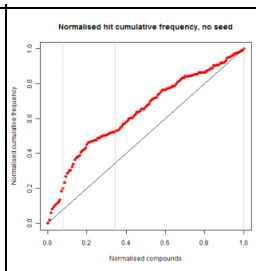
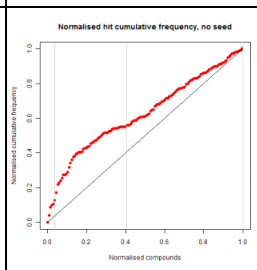
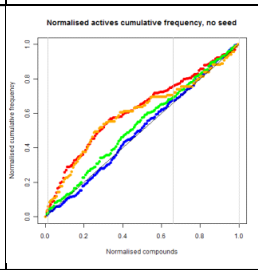
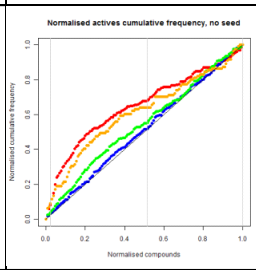
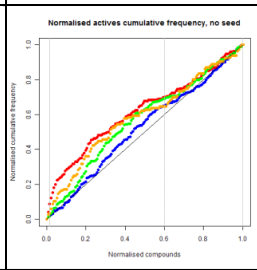
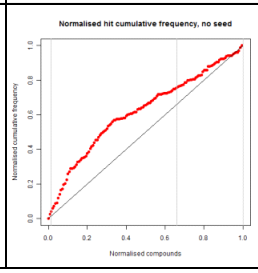
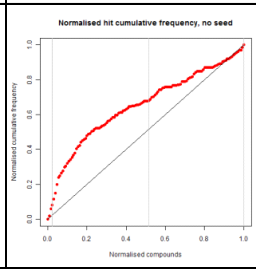
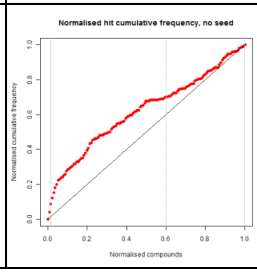
Algorithm		1	2	3
Active k optimisation	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

NMT-2 TcNMT

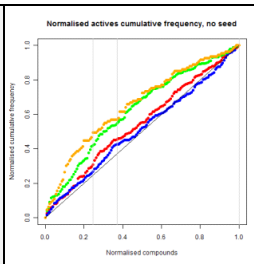
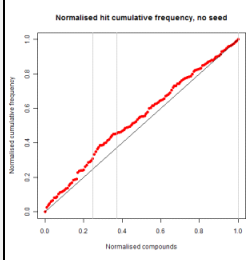
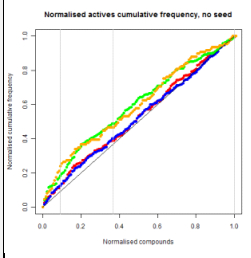
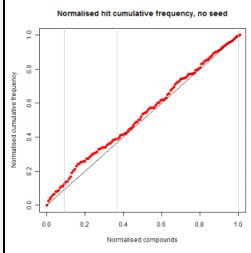
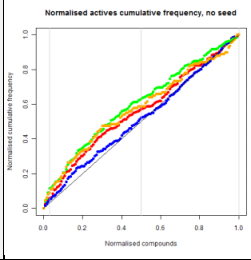
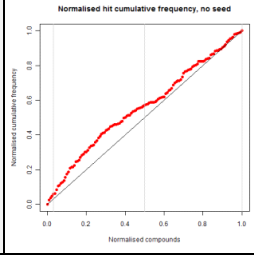
Algorithm		1	2	3
Active k optimisation (#1 might have used wrong start data - odd curves/deficiencies)	All actives			
	Hits			
SimplyGreedy	All actives			
	Hits			

B.2 **Preclustering learning curves**

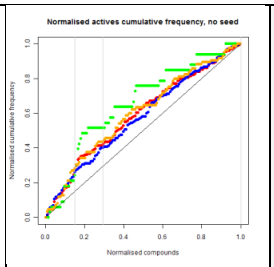
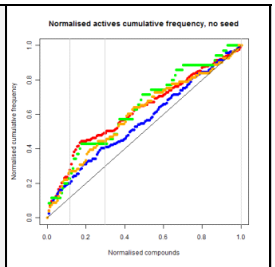
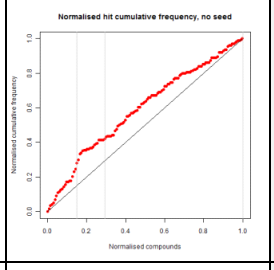
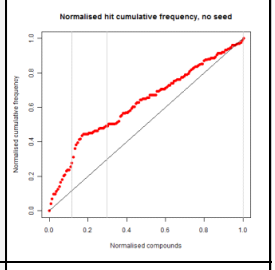
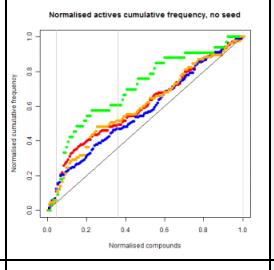
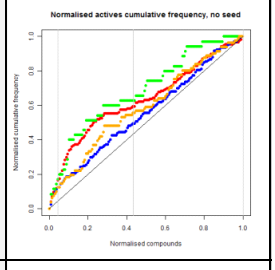
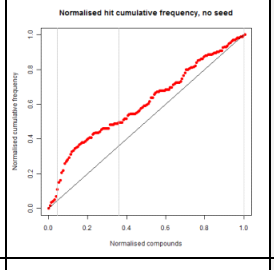
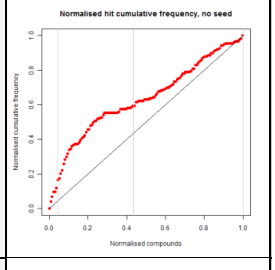
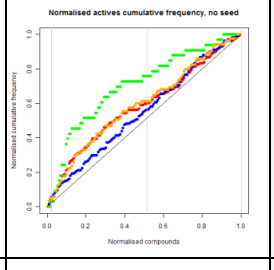
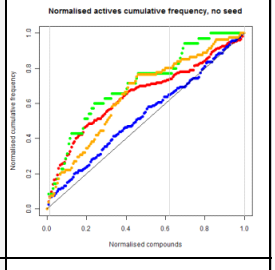
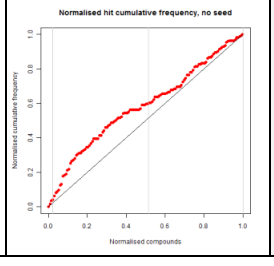
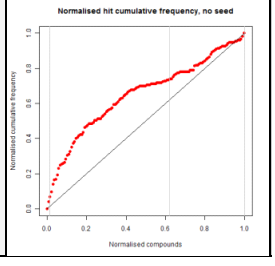
TS3 PvDHFR

Preclustering, TS>0.40	All actives			
	Hits			
Preclustering, TS>0.45	All actives			
	Hits			
Preclustering, TS>0.50	All actives			
	Hits			

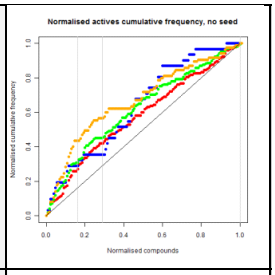
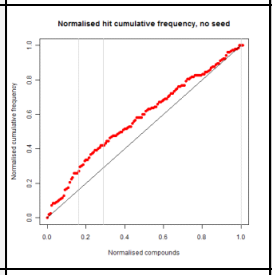
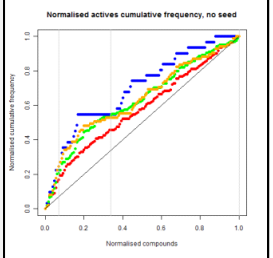
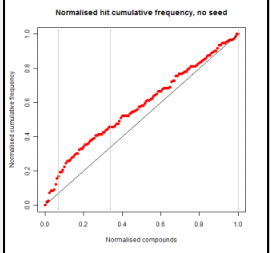
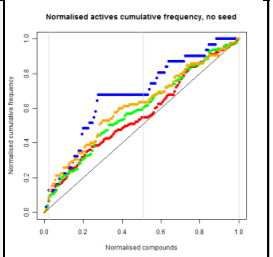
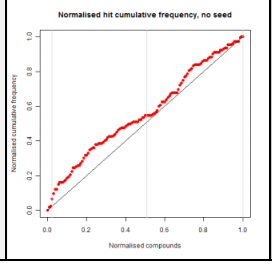
TS3 PfRdhfr

Preclustering, TS>0.40	All actives			
	Hits			
Preclustering, TS>0.45	All actives			
	Hits			
Preclustering, TS>0.50	All actives			
	Hits			

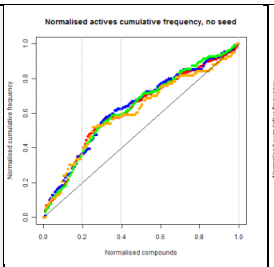
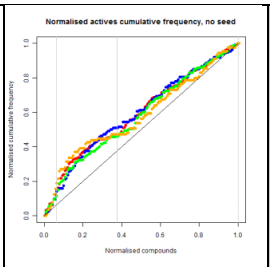
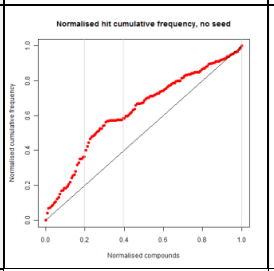
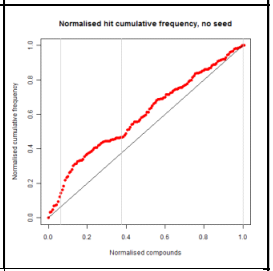
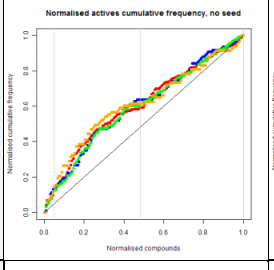
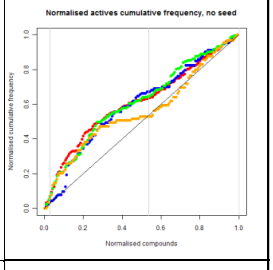
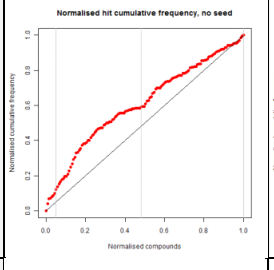
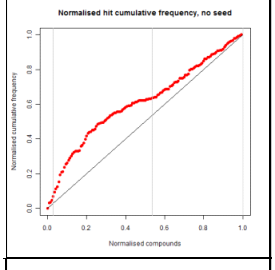
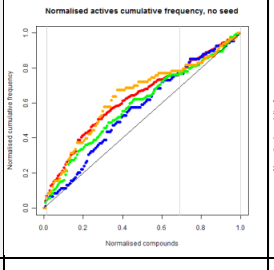
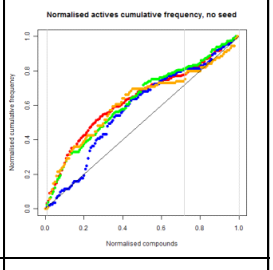
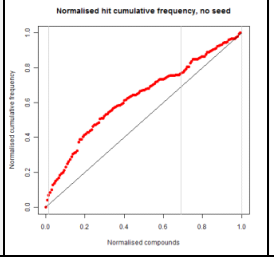
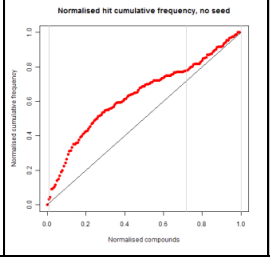
TS5 TcDHFR

Preclustering, TS>0.40	All actives			
	Hits			
Preclustering, TS>0.45	All actives			
	Hits			
Preclustering, TS>0.50	All actives			
	Hits			

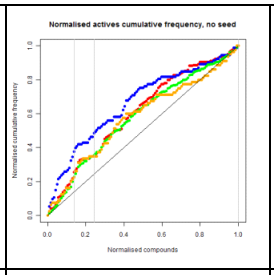
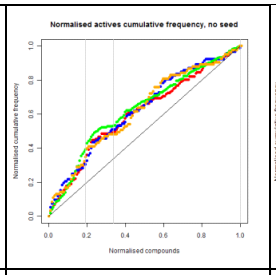
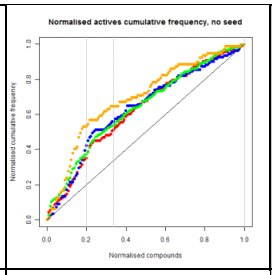
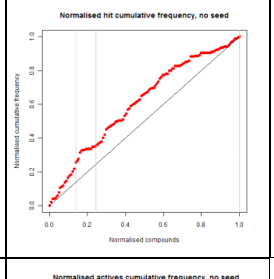
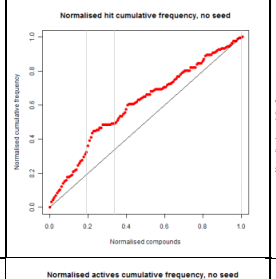
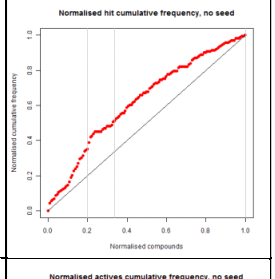
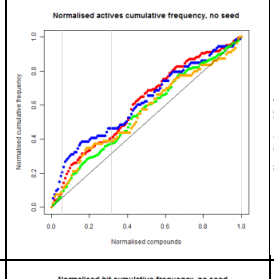
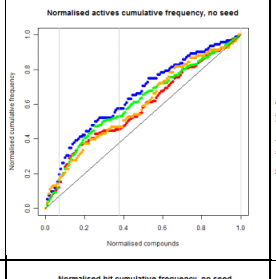
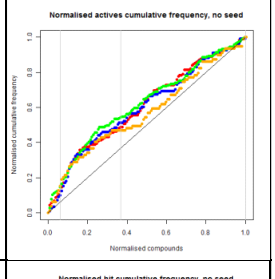
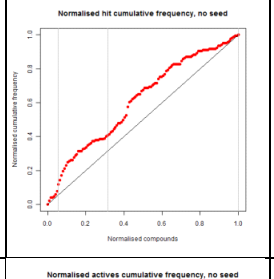
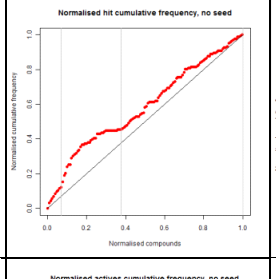
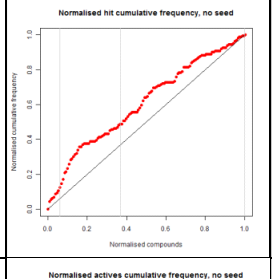
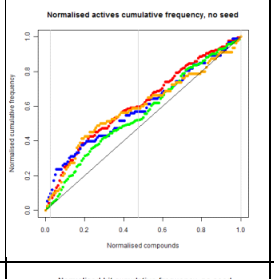
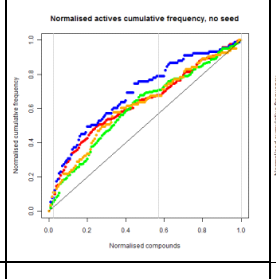
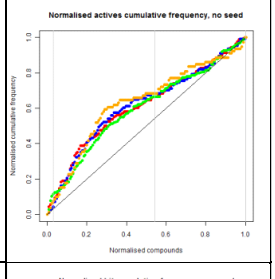
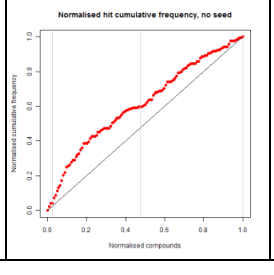
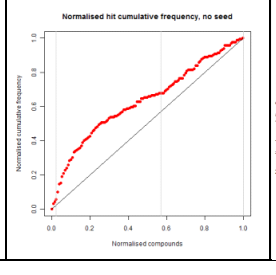
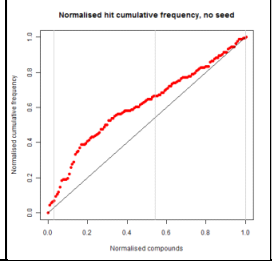
TS5 PfRdhfr

Preclustering, TS>0.40	All actives			
	Hits			
Preclustering, TS>0.45	All actives			
	Hits			
Preclustering, TS>0.50	All actives			
	Hits			

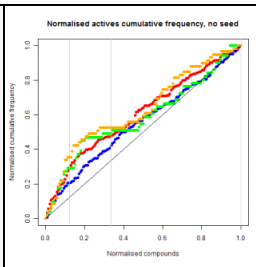
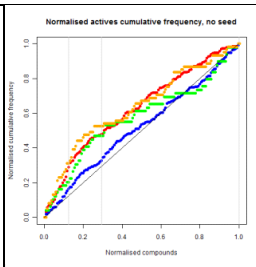
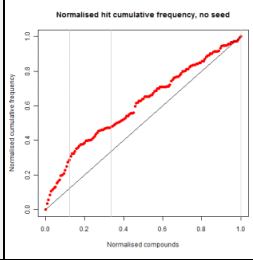
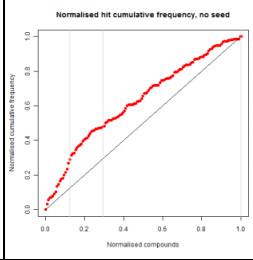
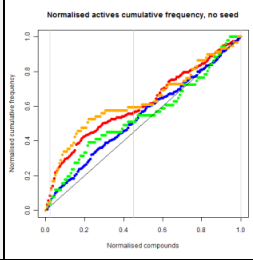
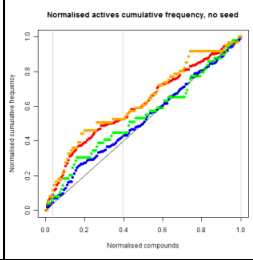
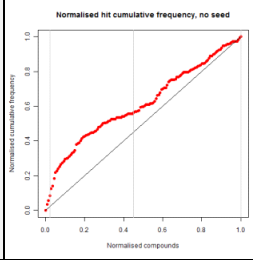
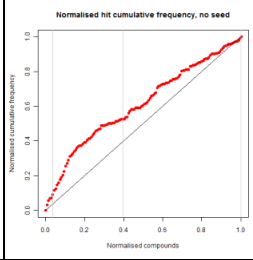
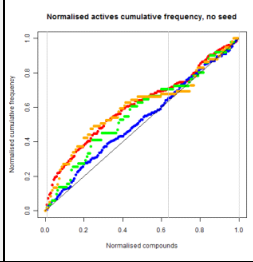
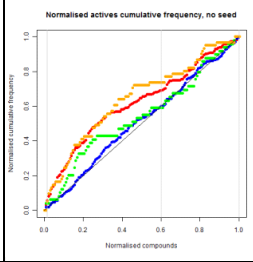
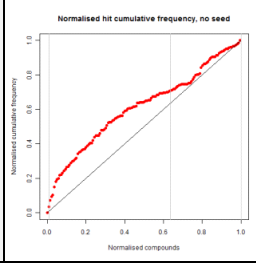
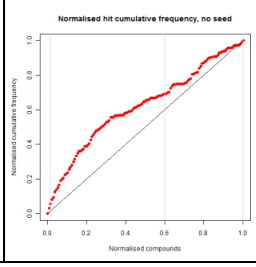
TS6 PvDHFR (8 other examples)

Preclustering, TS>0.40	All actives			
	Hits			
Preclustering, TS>0.45	All actives			
	Hits			
Preclustering, TS>0.50	All actives			
	Hits			

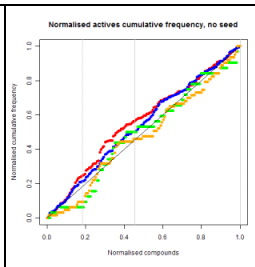
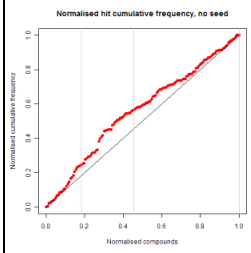
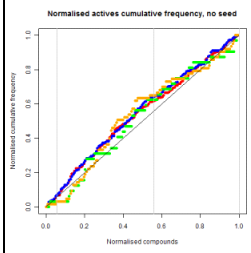
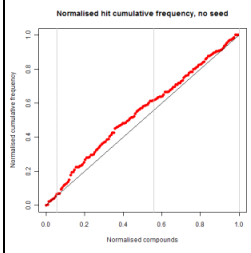
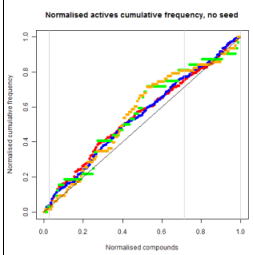
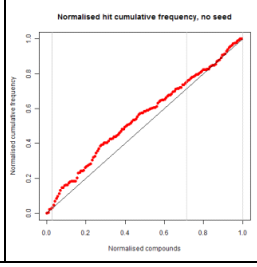
TS6 PfDHFR

Preclustering, TS>0.40	All actives			
	Hits			
Preclustering, TS>0.45	All actives			
	Hits			
Preclustering, TS>0.50	All actives			
	Hits			

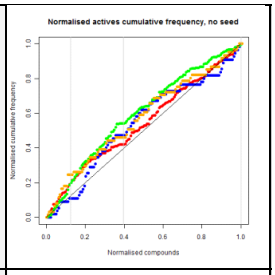
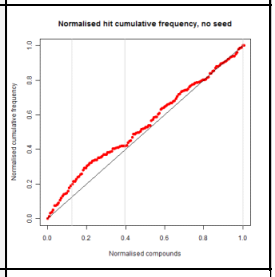
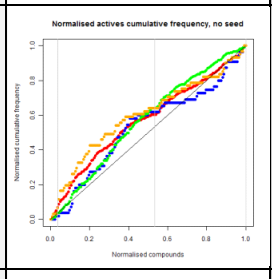
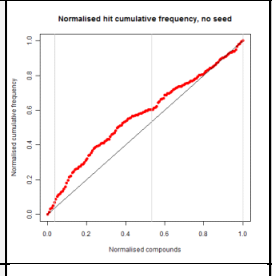
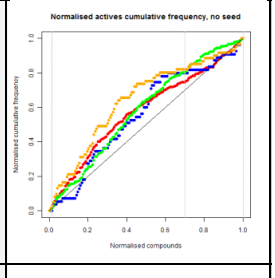
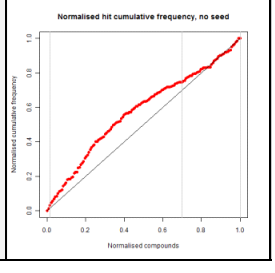
TS7 PvRdhfr

Preclustering, TS>0.40	All actives			
	Hits			
Preclustering, TS>0.45	All actives			
	Hits			
Preclustering, TS>0.50	All actives			
	Hits			

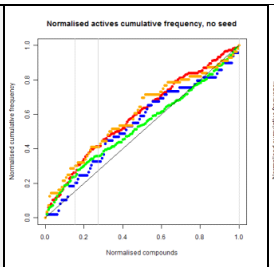
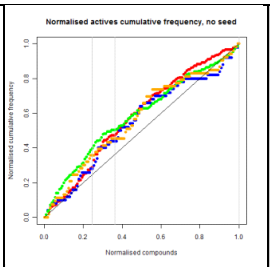
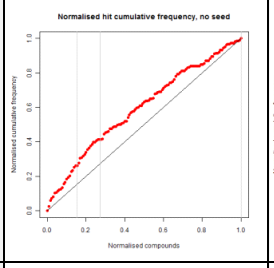
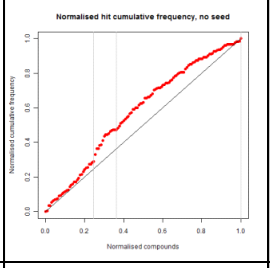
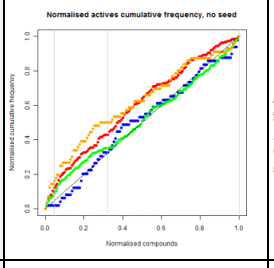
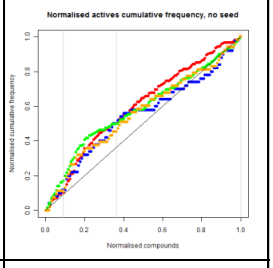
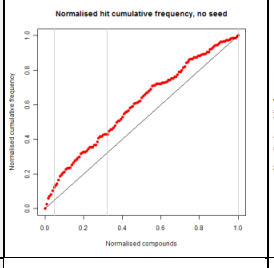
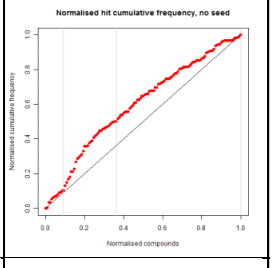
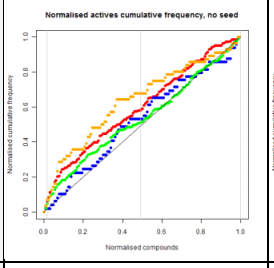
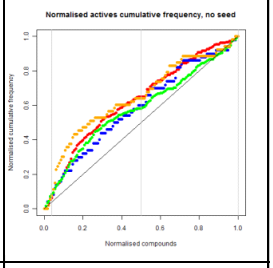
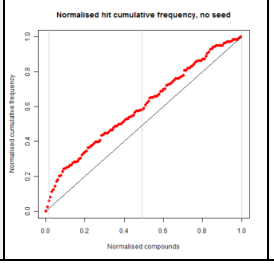
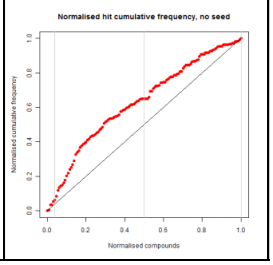
PGK-2 TcPGK

Preclustering, TS>0.40	All actives			
	Hits			
Preclustering, TS>0.45	All actives			
	Hits			
Preclustering, TS>0.50	All actives			
	Hits			

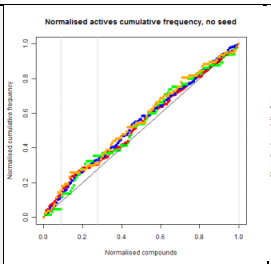
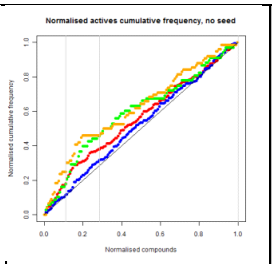
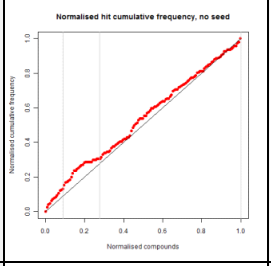
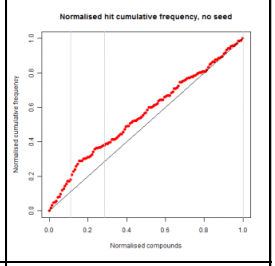
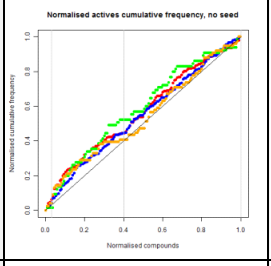
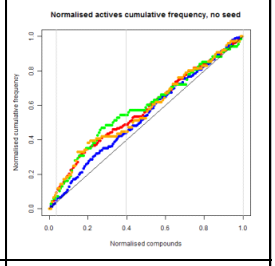
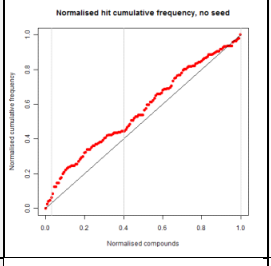
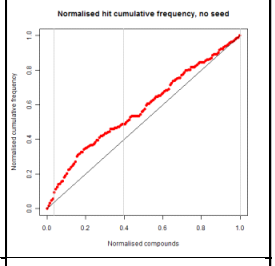
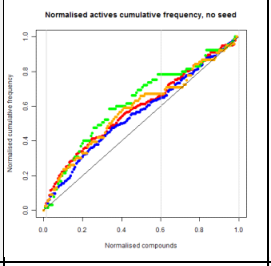
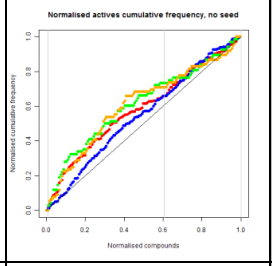
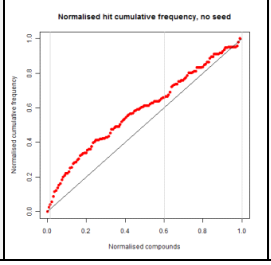
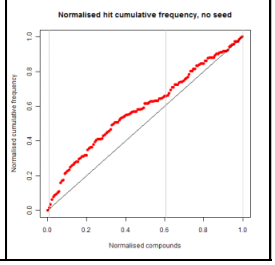
NMT-1 TbNMT

Preclustering, TS>0.40	All actives			
	Hits			
Preclustering, TS>0.45	All actives			
	Hits			
Preclustering, TS>0.50	All actives			
	Hits			

NMT-1 PvNMT

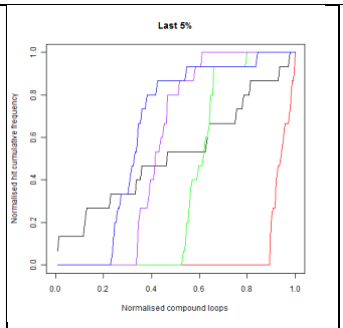
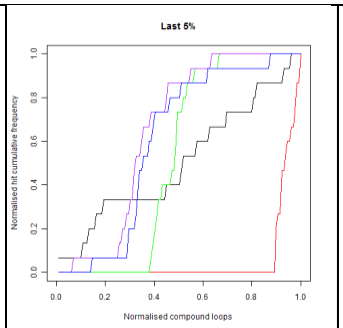
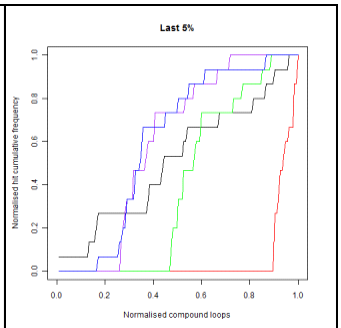
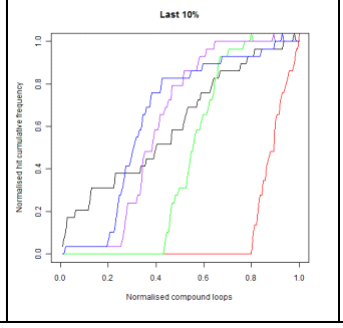
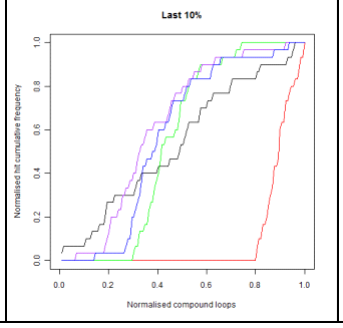
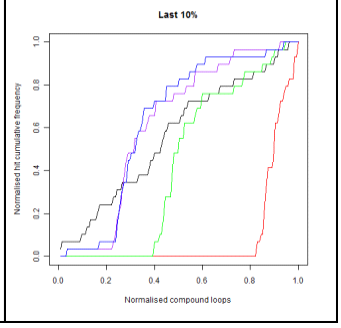
Preclustering, TS>0.40	All actives			
	Hits			
Preclustering, TS>0.45	All actives			
	Hits			
Preclustering, TS>0.50	All actives			
	Hits			

NMT-2 TcNMT

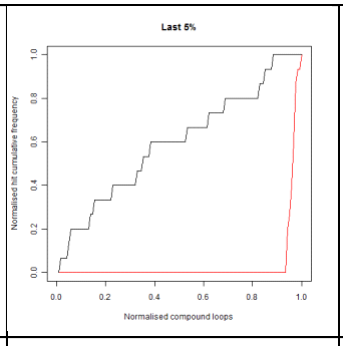
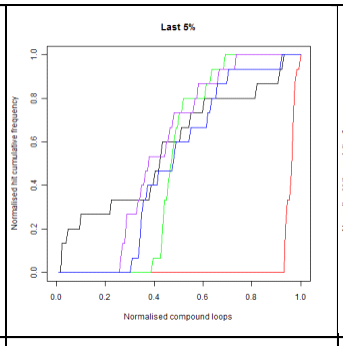
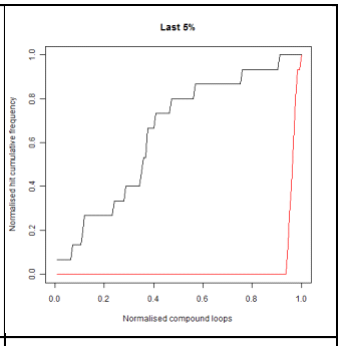
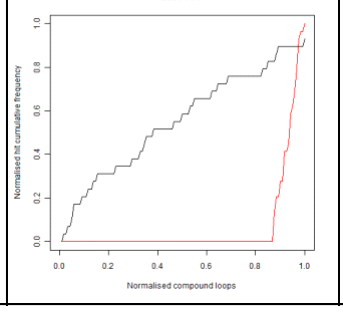
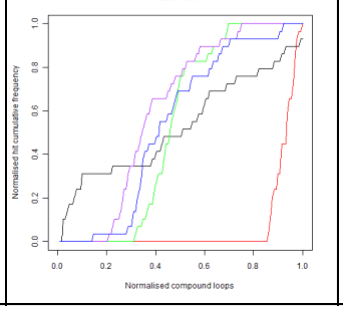
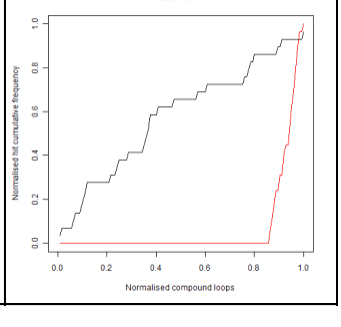
Preclustering, TS>0.40	All actives			
	Hits			
Preclustering, TS>0.45	All actives			
	Hits			
Preclustering, TS>0.50	All actives			
	Hits			

B.3 Rare category detection for active k-optimisation and SimplyGreedy

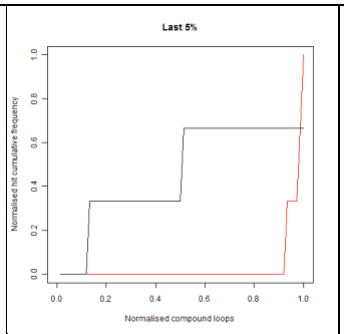
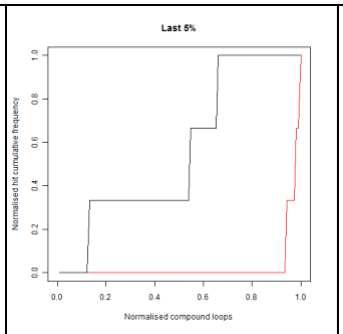
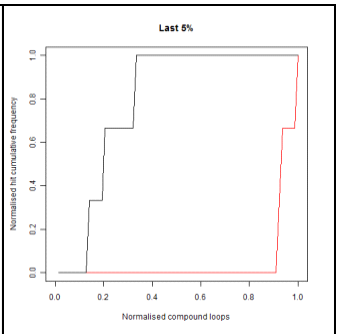
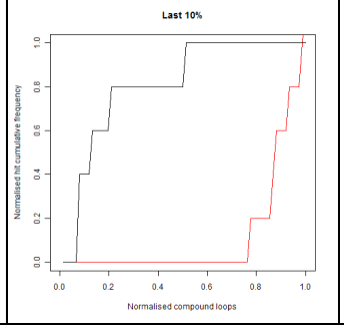
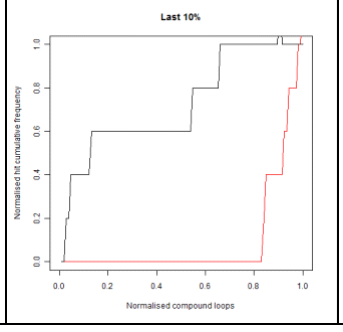
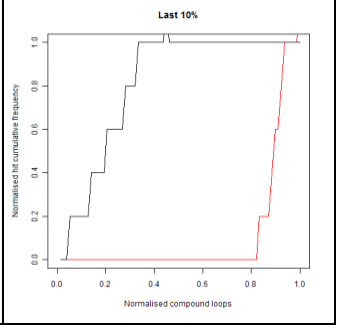
TS3 PvDHFR

Last 5%			
Last 10%			

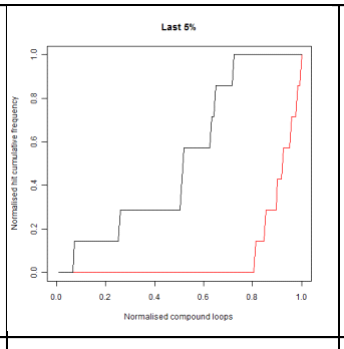
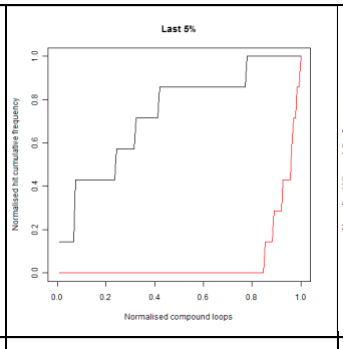
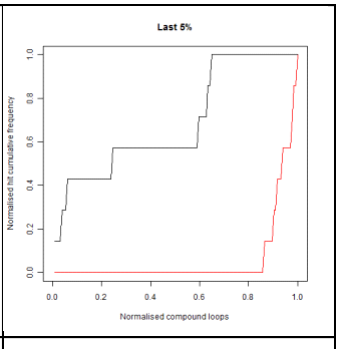
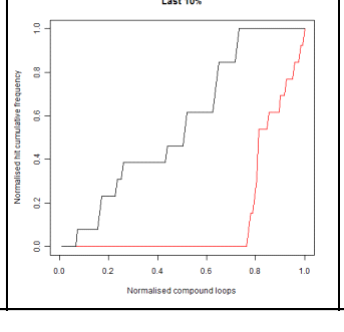
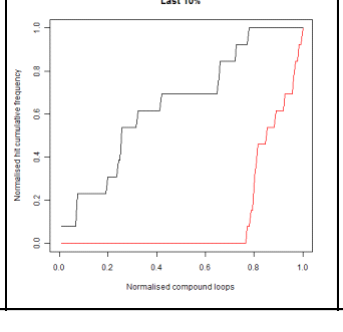
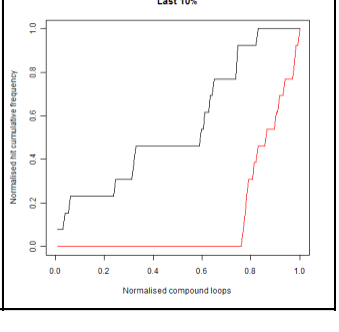
TS3 PfRdhfr

Last 5%			
Last 10%			

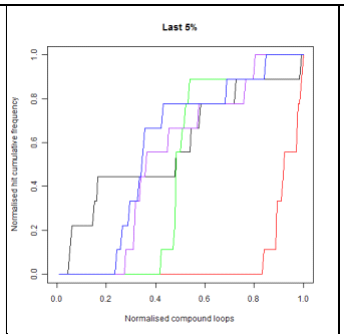
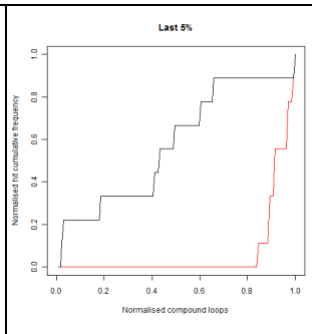
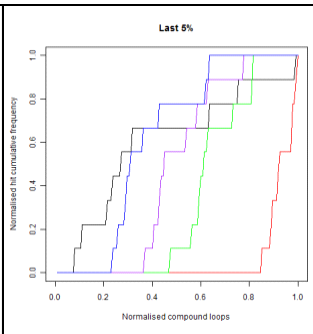
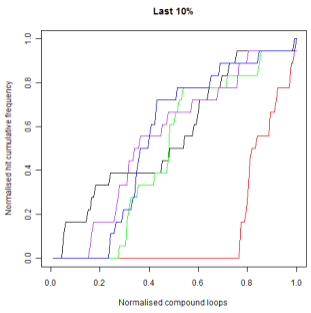
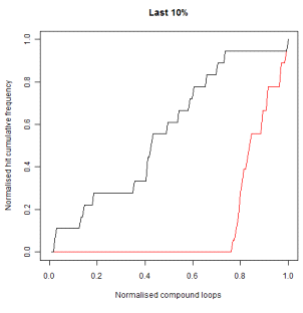
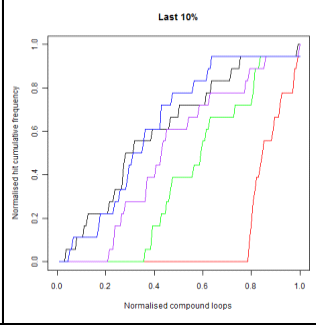
TS4 TbDHFR

Last 5%			
Last 10%			

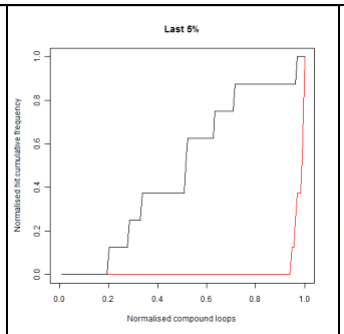
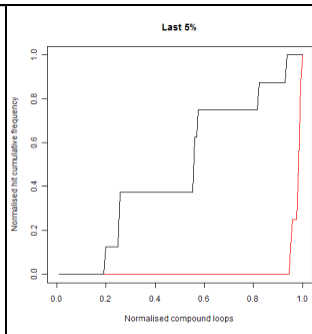
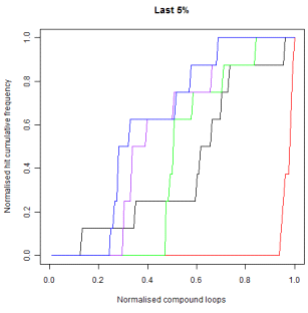
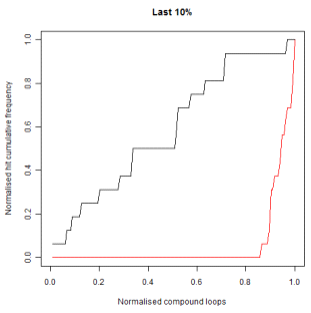
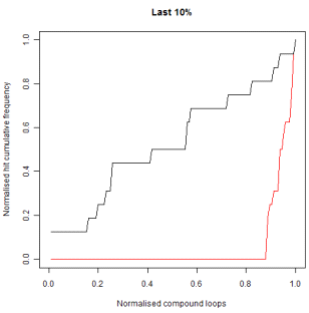
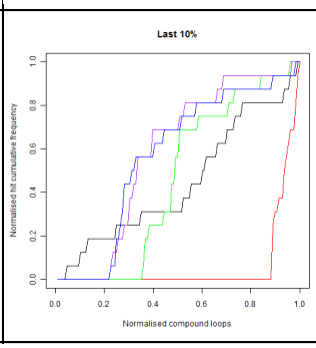
TS4 SmDHFR

Last 5%			
Last 10%			

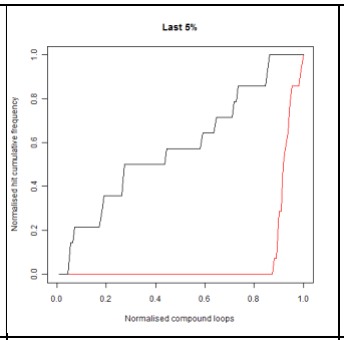
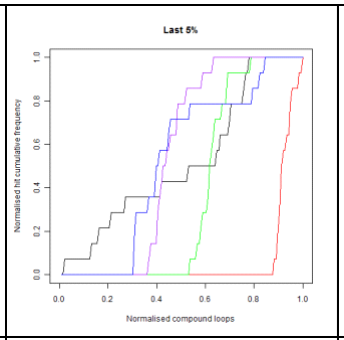
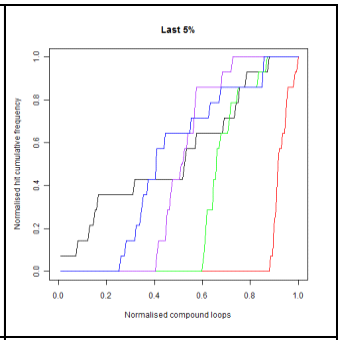
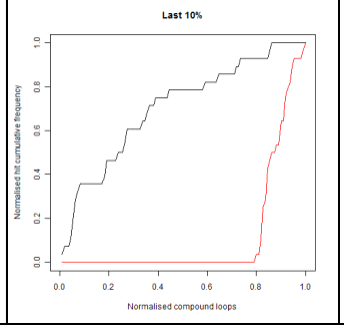
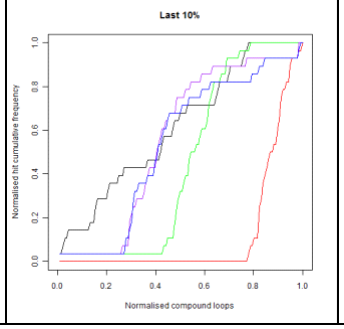
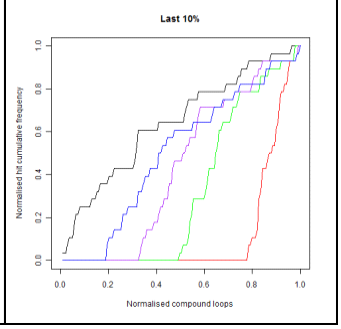
TS5 TcDHFR

Last 5%			
Last 10%			

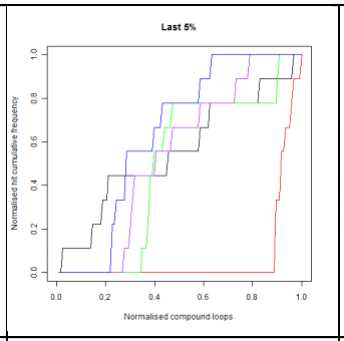
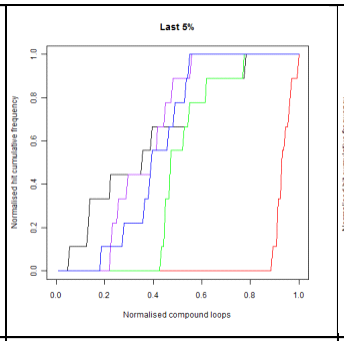
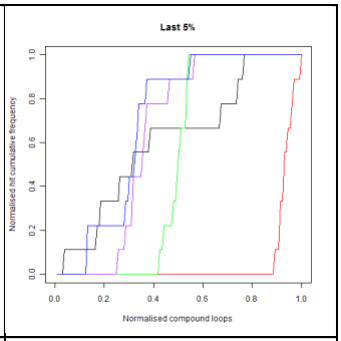
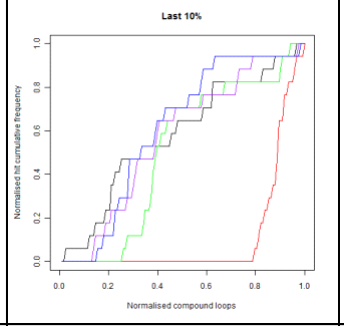
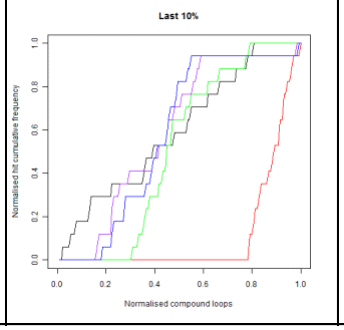
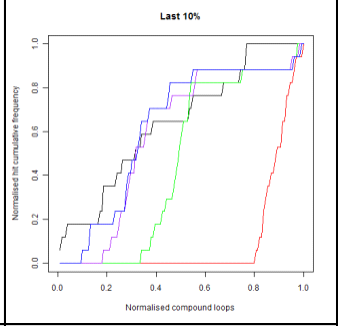
TS5 PvRdhfr

Last 5%			
Last 10%			

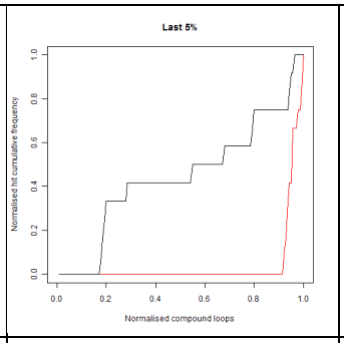
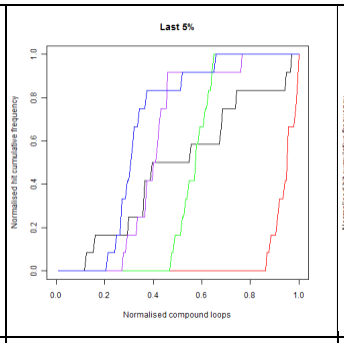
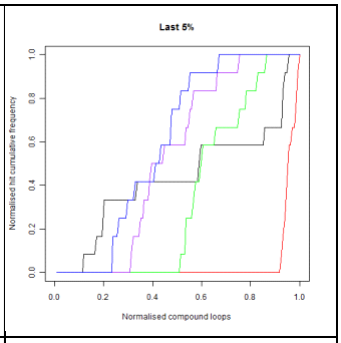
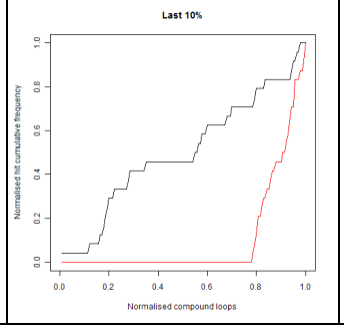
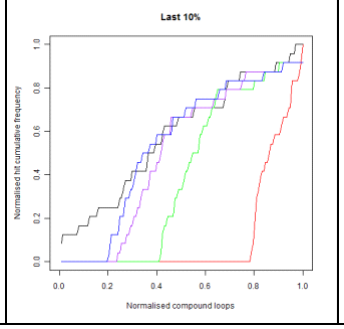
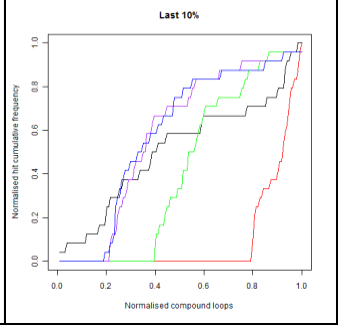
TS6 PvDHFR

Last 5%			
Last 10%			

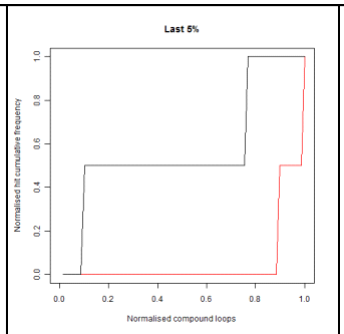
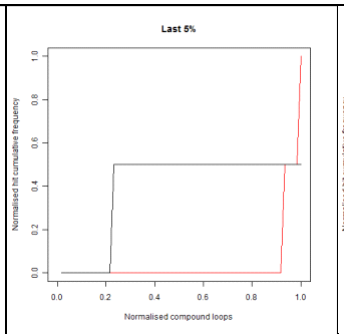
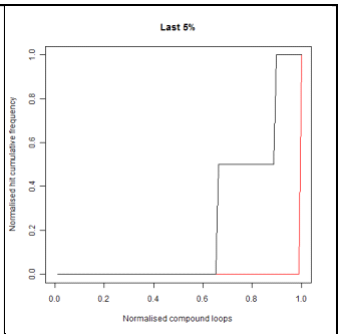
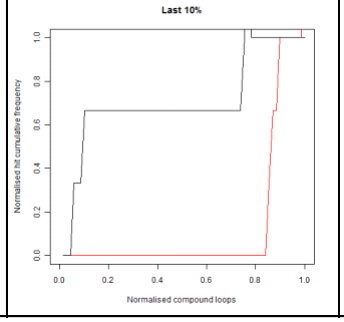
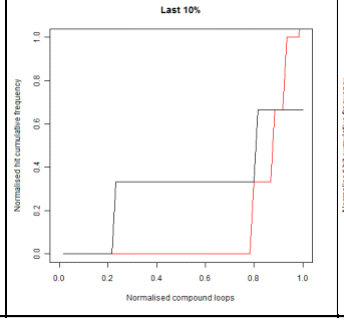
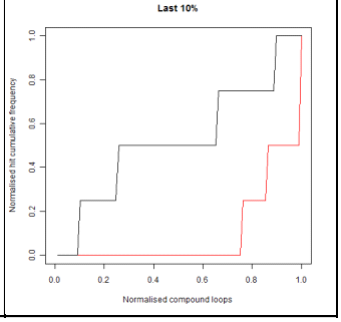
TS6 PfDHFR

Last 5%			
Last 10%			

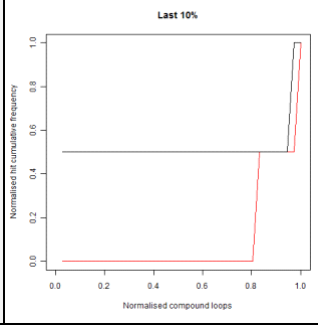
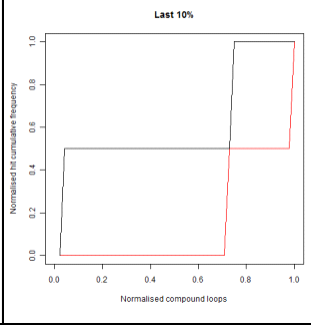
TS7 PvRdhfr

Last 5%			
Last 10%			

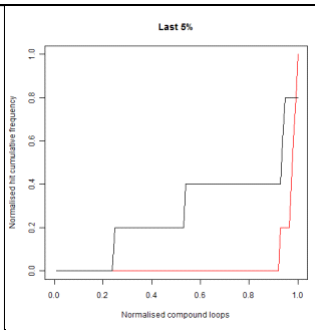
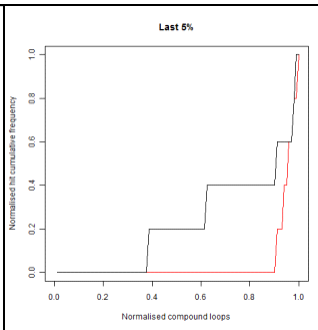
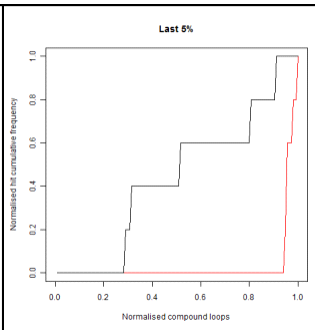
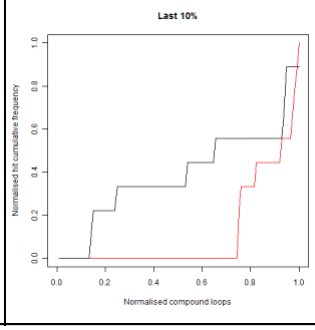
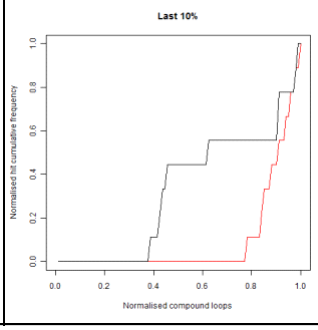
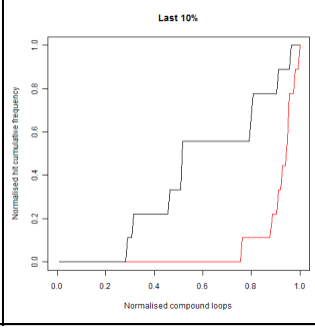
TS7 LmDHFR

Last 5%			
Last 10%			

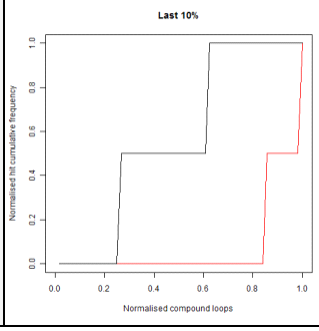
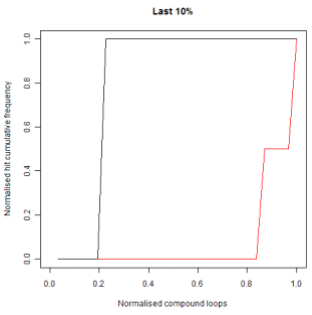
PGK-1 TbPGK

Last 5%			
Last 10%			

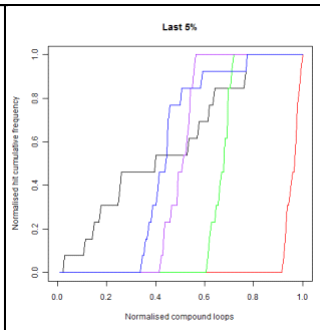
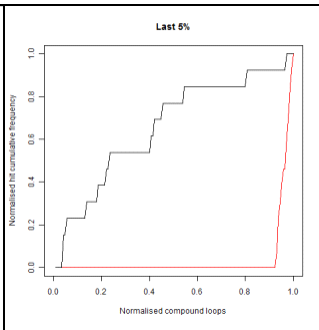
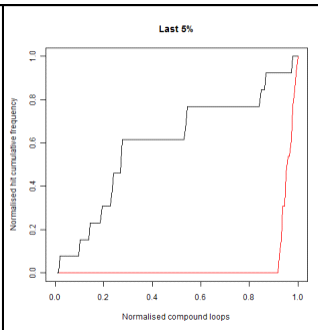
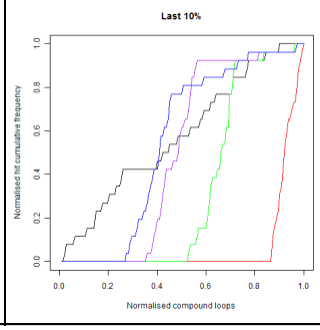
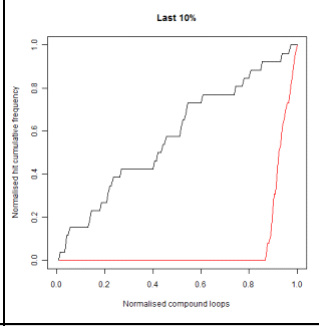
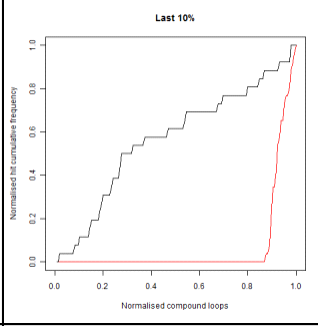
PGK-1 PvPGK

Last 5%			
Last 10%			

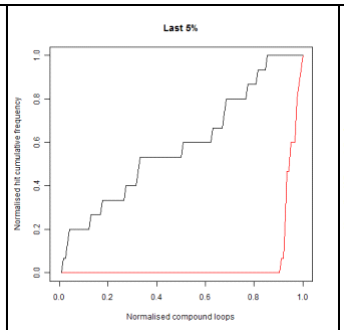
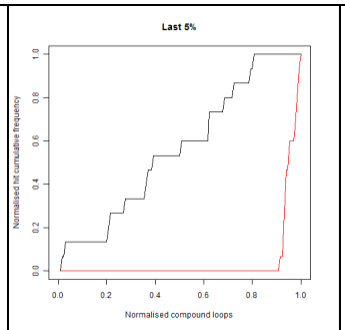
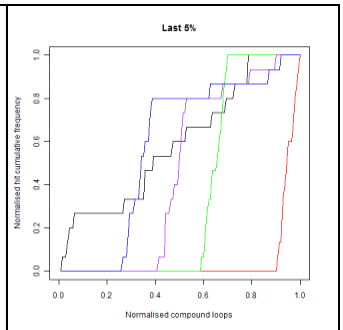
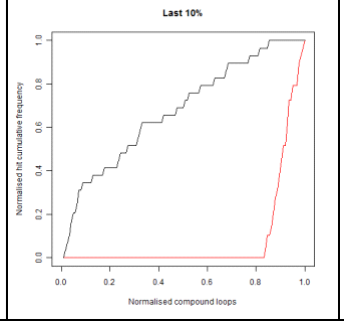
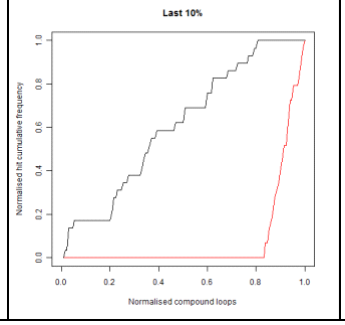
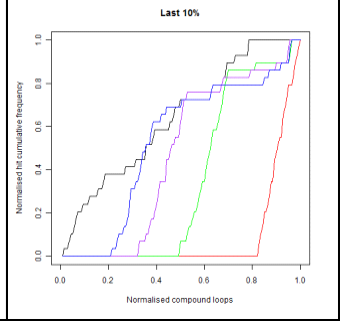
PGK-2 SmPGK

Last 5%			
Last 10%			

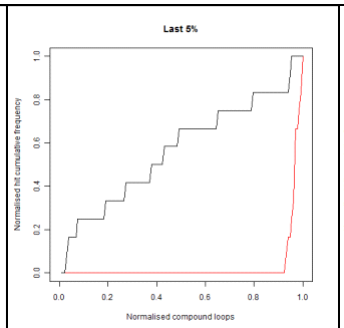
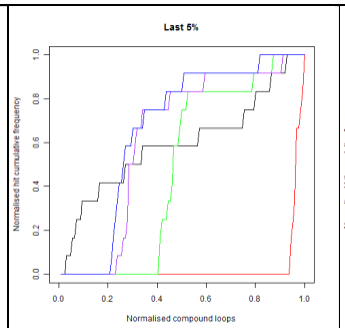
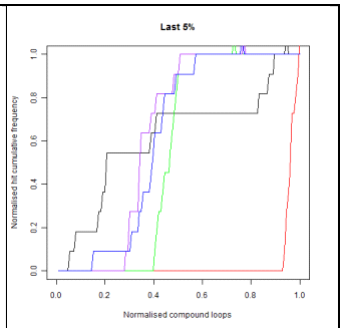
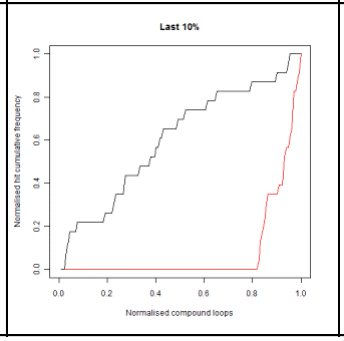
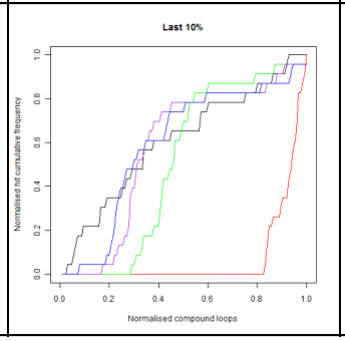
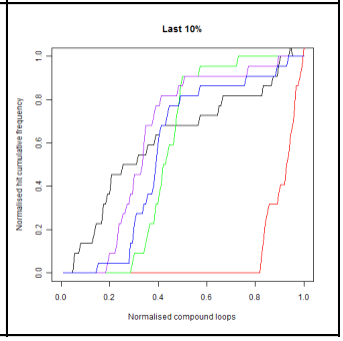
PGK-2 TcPGK

Last 5%			
Last 10%			

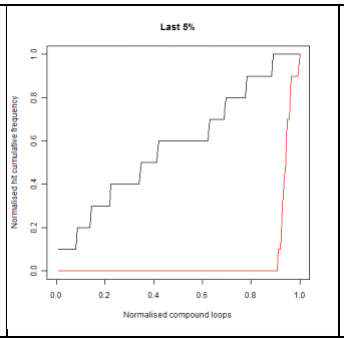
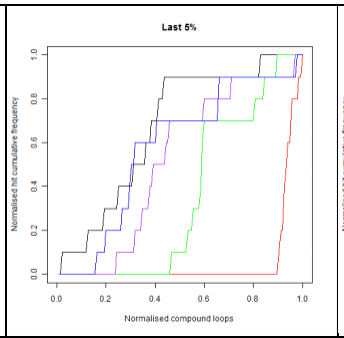
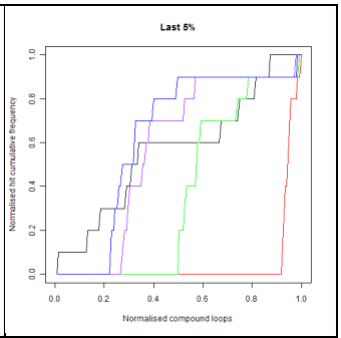
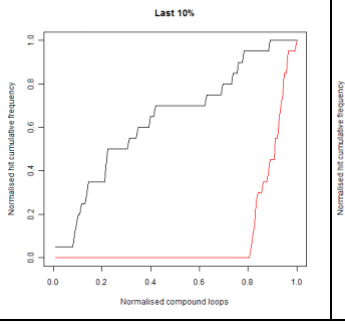
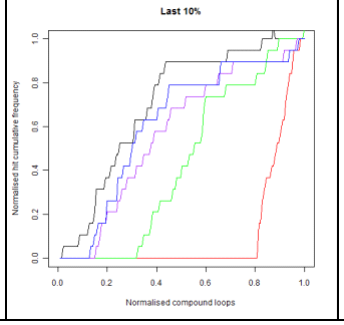
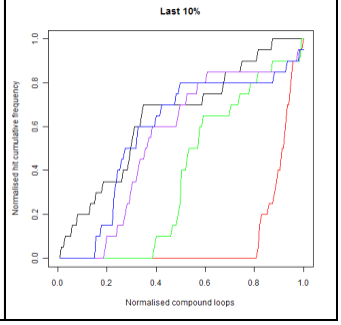
NMT-1 TbNMT

Last 5%			
Last 10%			

NMT-1 PvNMT

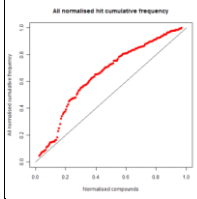
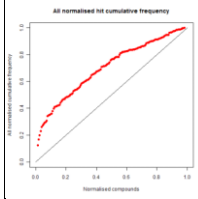
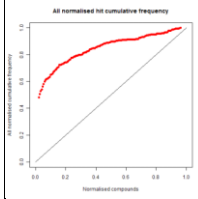
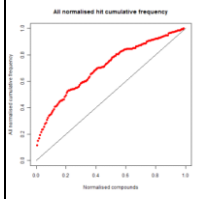
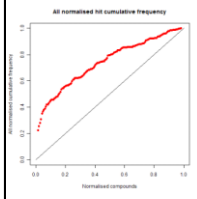
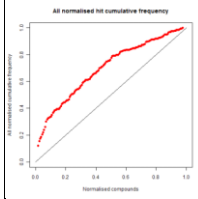
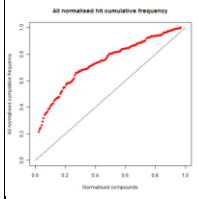
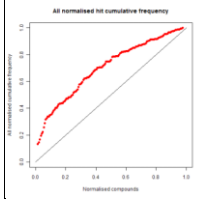
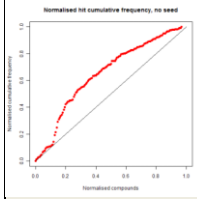
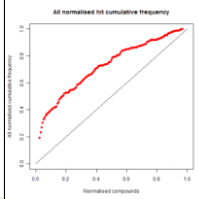
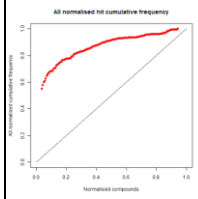
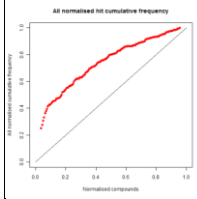
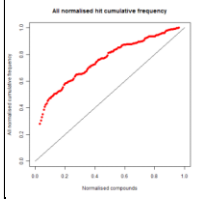
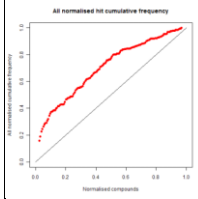
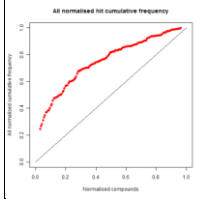
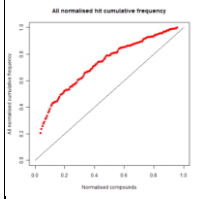
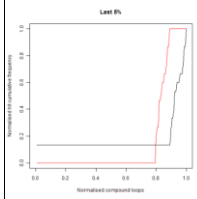
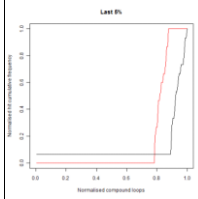
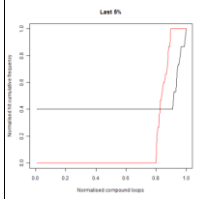
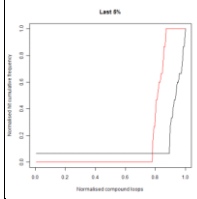
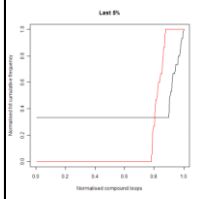
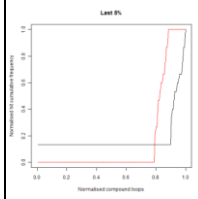
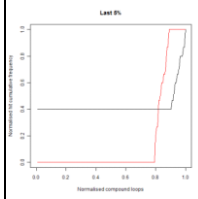
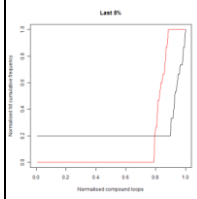
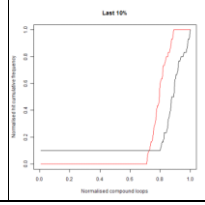
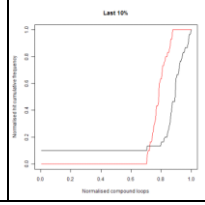
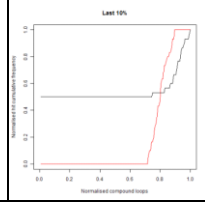
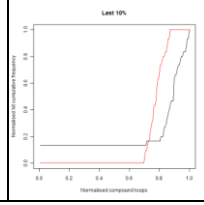
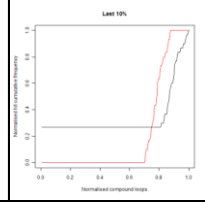
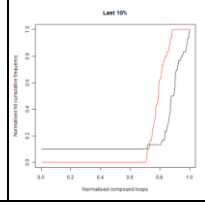
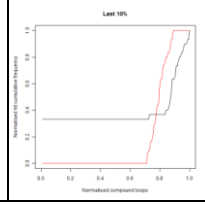
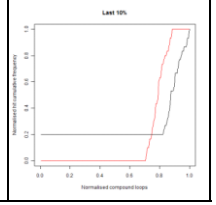
Last 5%			
Last 10%			

NMT-2 TcNMT

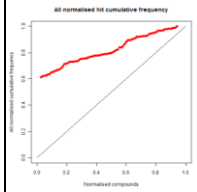
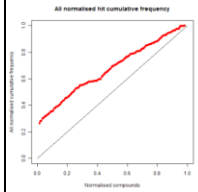
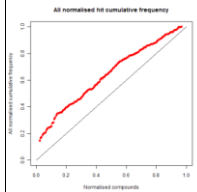
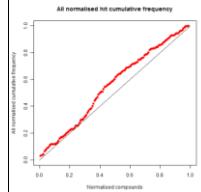
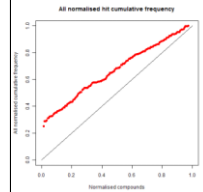
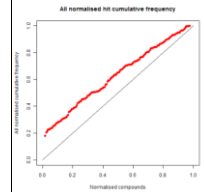
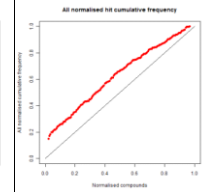
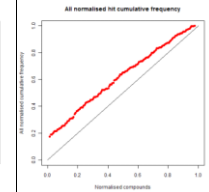
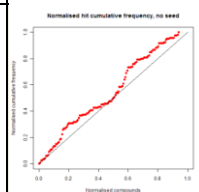
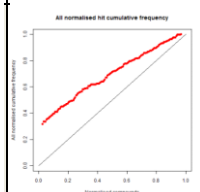
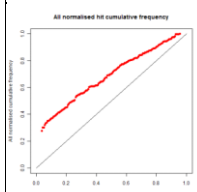
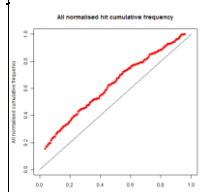
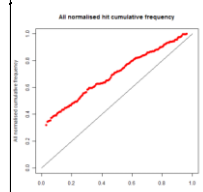
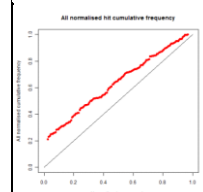
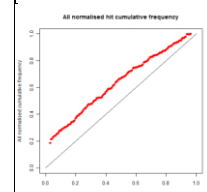
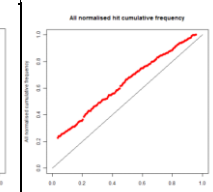
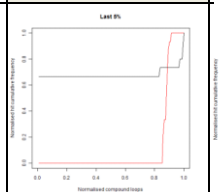
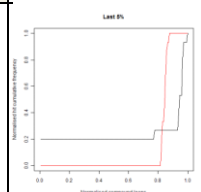
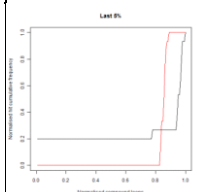
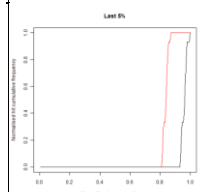
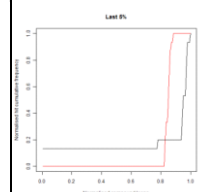
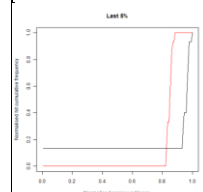
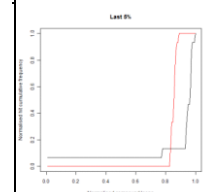
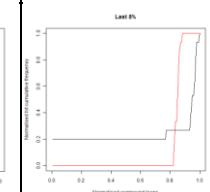
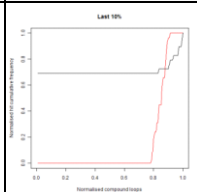
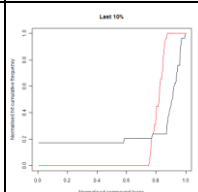
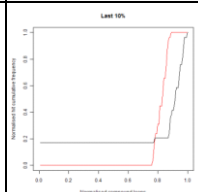
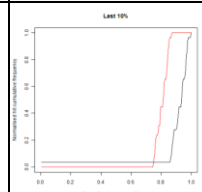
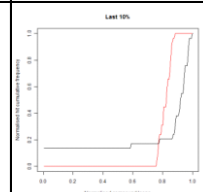
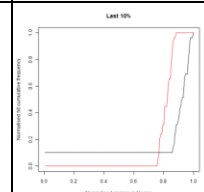
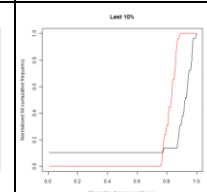
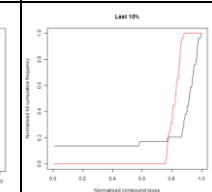
Last 5%			
Last 10%			

B.4 Transfer Learning, with rare category detection comparison versus the SimplyGreedy curves

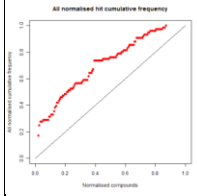
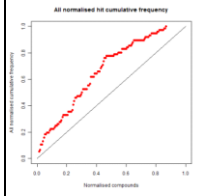
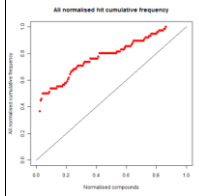
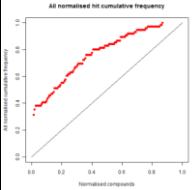
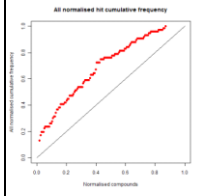
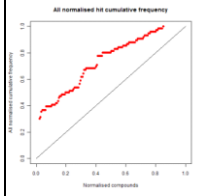
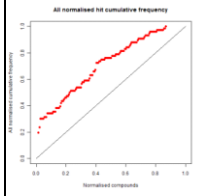
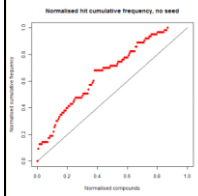
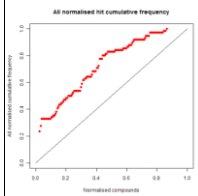
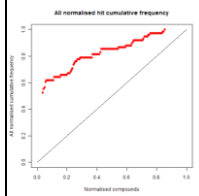
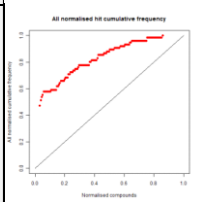
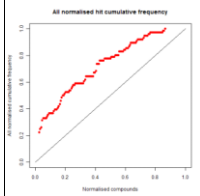
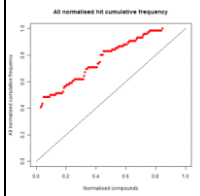
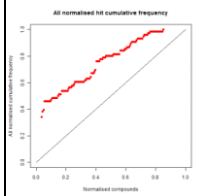
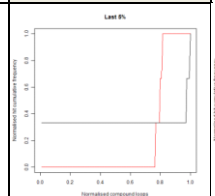
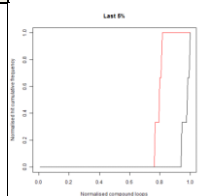
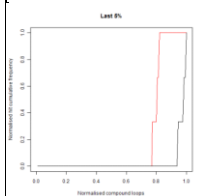
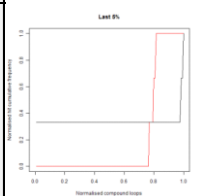
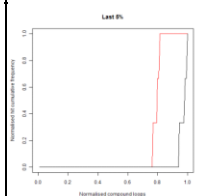
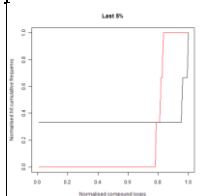
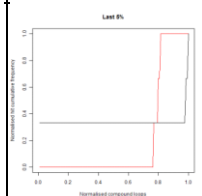
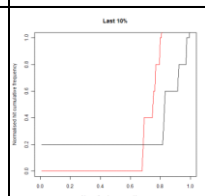
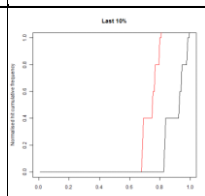
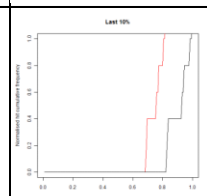
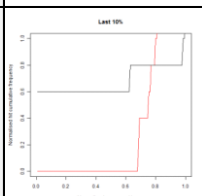
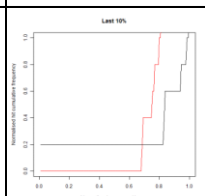
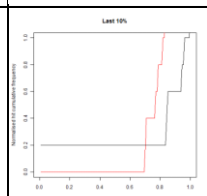
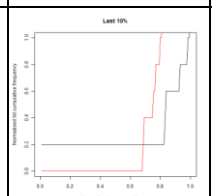
TS3 PvDHFR

	Hs internal seed	External seed parasite/protein						
	HsDHFR	PfDHFR	PvDHFR	TbDHFR	TcDHFR	PvNMT	TbNMT	TcNMT
Single seed, all hits								
Multi seed (Hs, toxic, parasite), all hits								
Last 5%								
Last 10%								

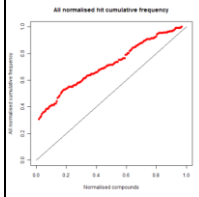
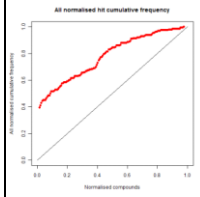
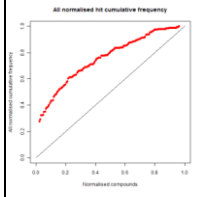
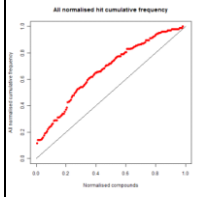
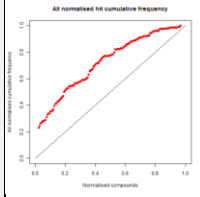
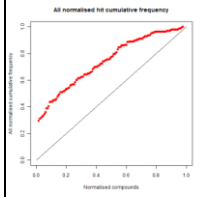
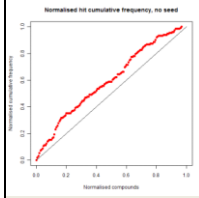
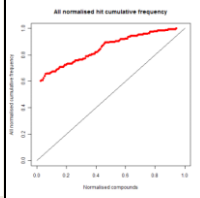
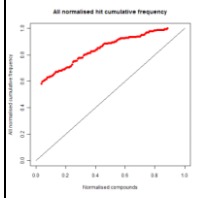
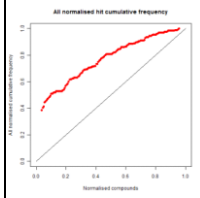
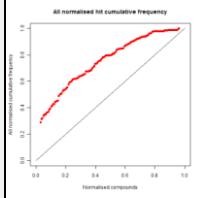
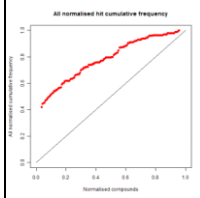
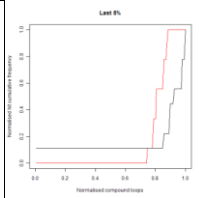
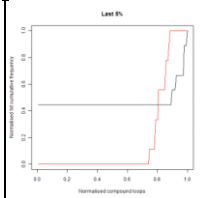
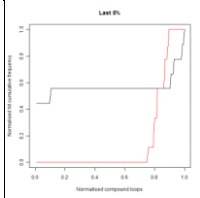
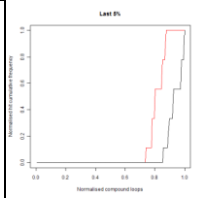
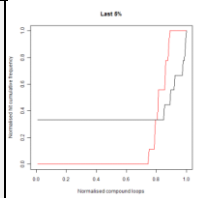
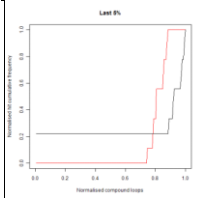
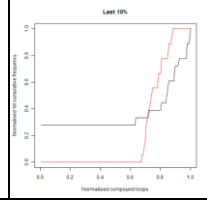
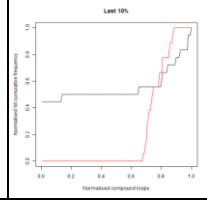
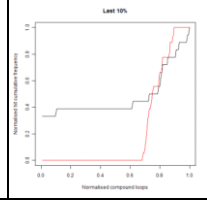
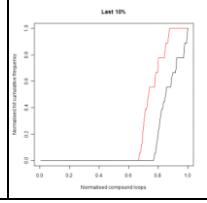
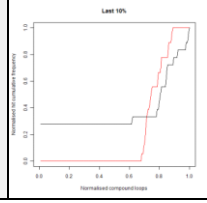
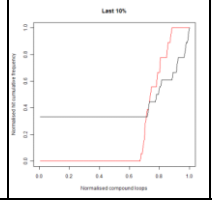
TS3 PfRdhfr

	Hs internal seed	External seed parasite/protein						
	HsDHFR	PfDHFR	PvDHFR	TbDHFR	TcDHFR	PvNMT	TbNMT	TcNMT
Single seed, all hits								
Multi seed (Hs, toxic, parasite), all hits								
Last 5%								
Last 10%								

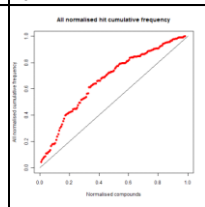
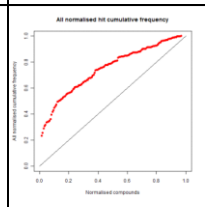
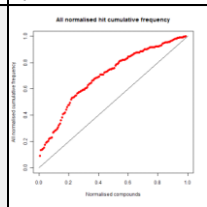
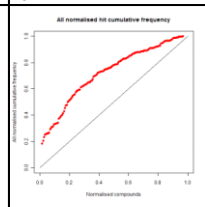
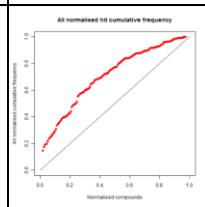
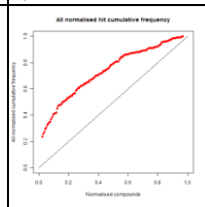
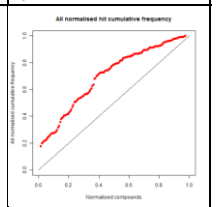
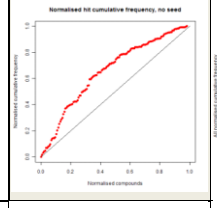
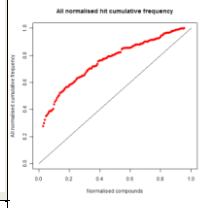
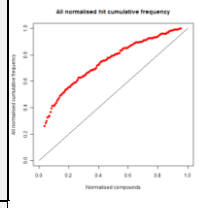
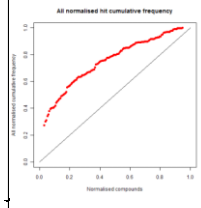
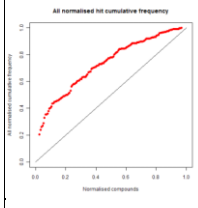
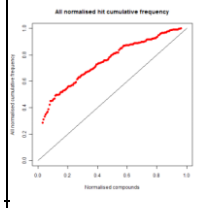
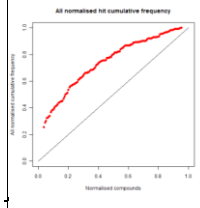
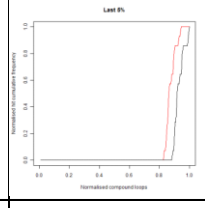
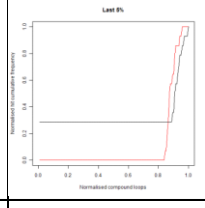
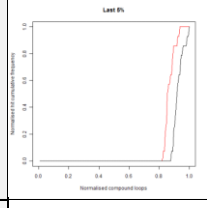
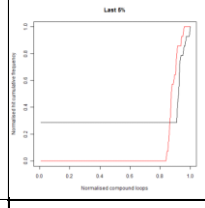
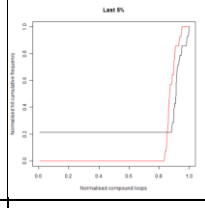
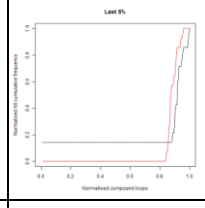
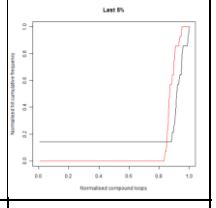
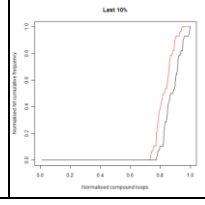
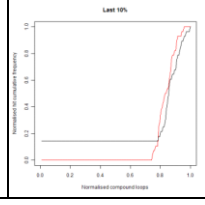
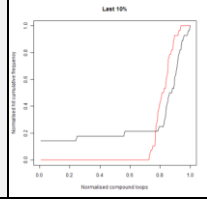
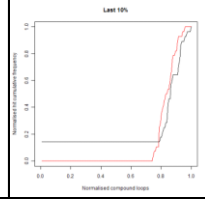
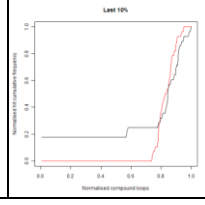
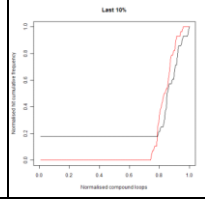
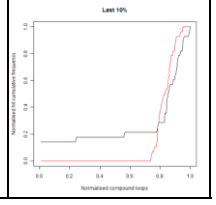
TS4 TbDHFR

	Hs internal seed	External seed parasite/protein						
	HsDHFR	PfDHFR	PvDHFR	TbDHFR	TcDHFR	PvNMT	TbNMT	TcNMT
Single seed, all hits								
Multi seed (Hs, toxic, parasite), all hits								
Last 5%								
Last 10%								

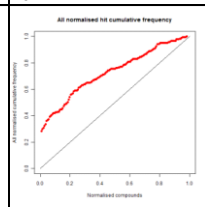
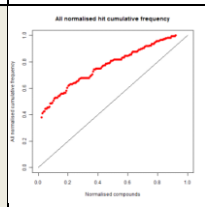
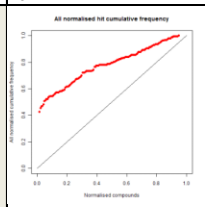
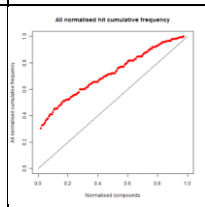
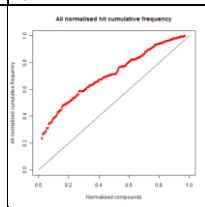
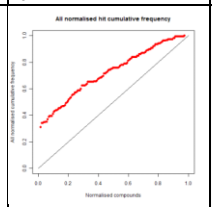
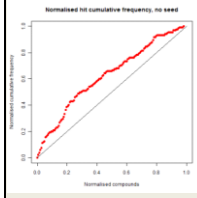
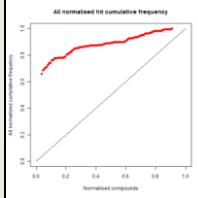
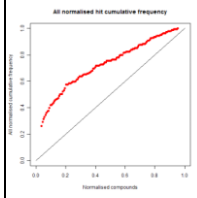
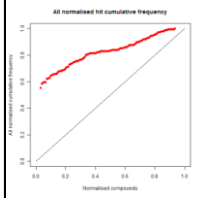
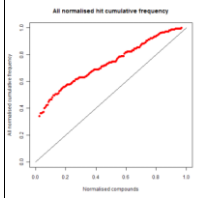
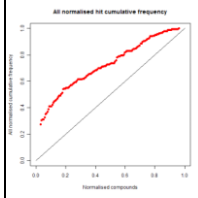
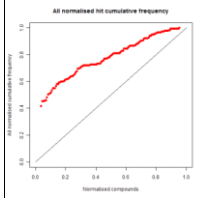
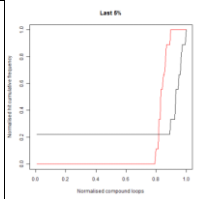
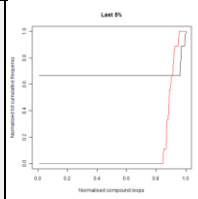
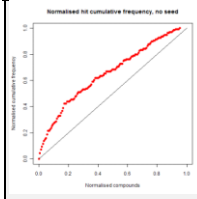
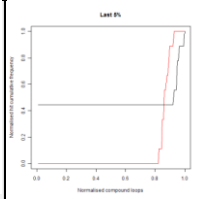
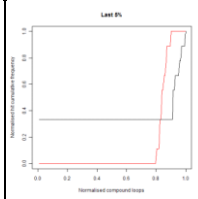
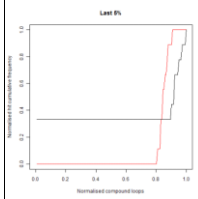
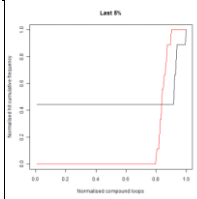
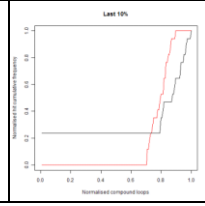
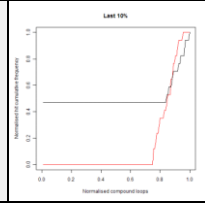
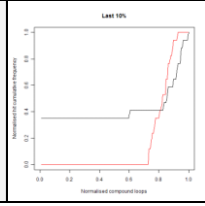
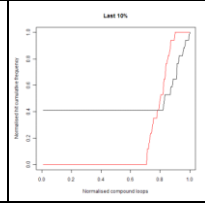
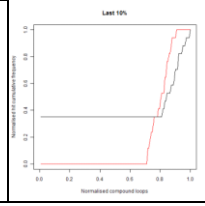
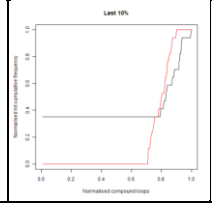
TS5 TcDHFR

	Hs internal seed	External seed parasite/protein						
	HsDHFR	PfDHFR	PvDHFR	TbDHFR	TcDHFR	PvNMT	TbNMT	TcNMT
Single seed, all hits								
Multi seed (Hs, toxic, parasite), all hits								
Last 5%								
Last 10%								

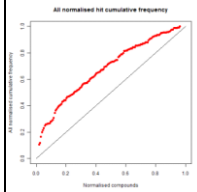
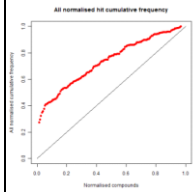
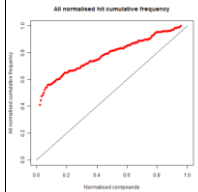
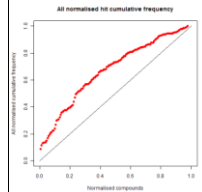
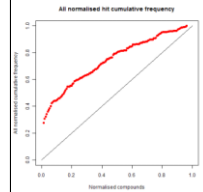
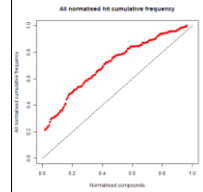
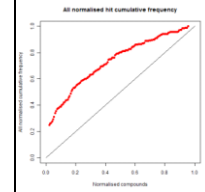
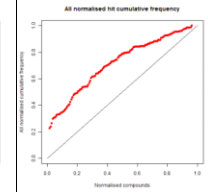
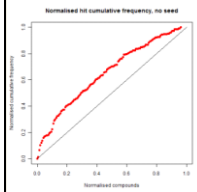
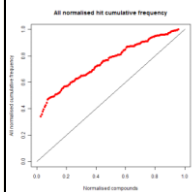
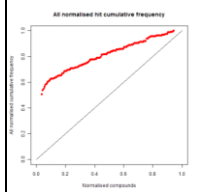
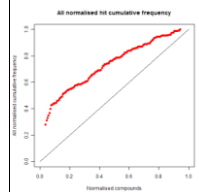
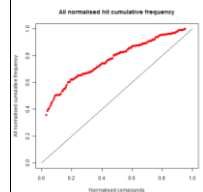
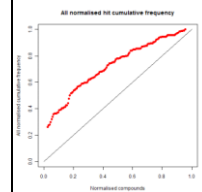
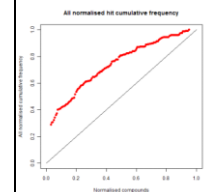
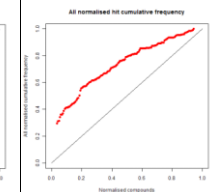
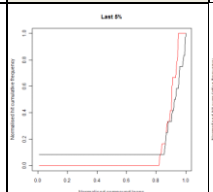
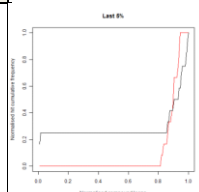
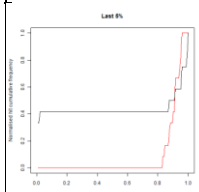
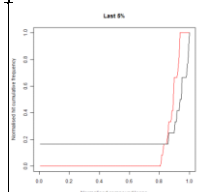
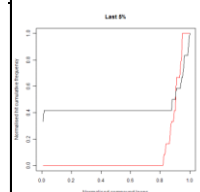
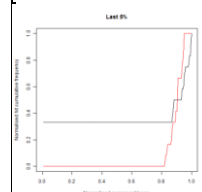
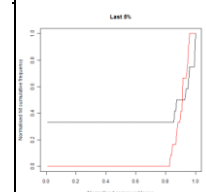
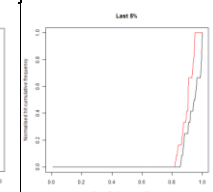
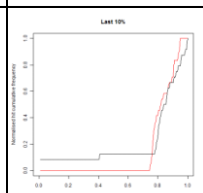
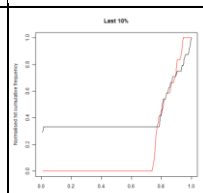
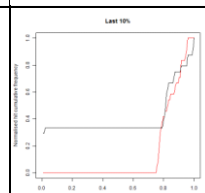
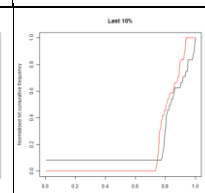
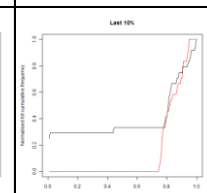
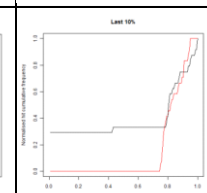
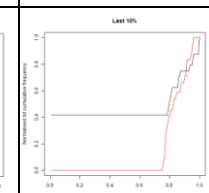
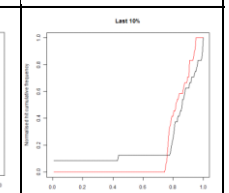
TS6 PvDHFR

	Hs internal seed	External seed parasite/protein						
	HsDHFR	PfDHFR	PvDHFR	TbDHFR	TcDHFR	PvNMT	TbNMT	TcNMT
Single seed, all hits								
Multi seed (Hs, toxic, parasite), all hits								
Last 5%								
Last 10%								

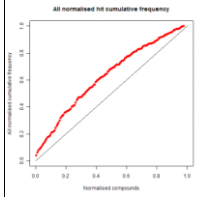
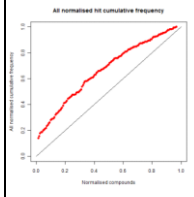
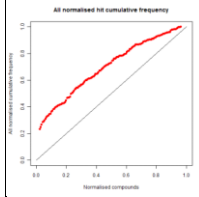
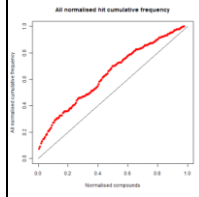
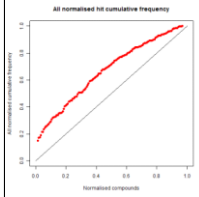
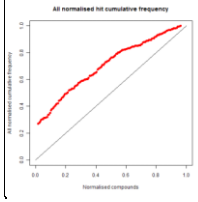
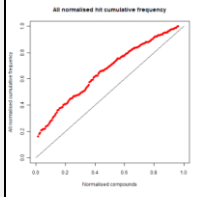
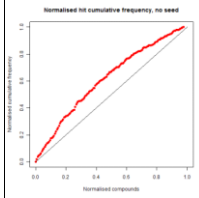
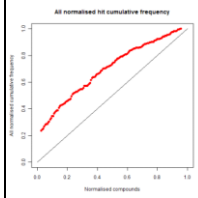
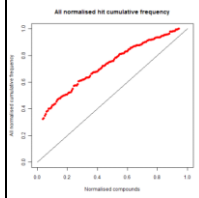
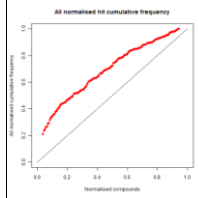
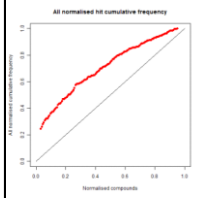
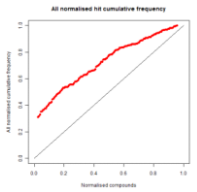
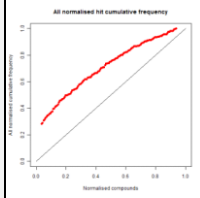
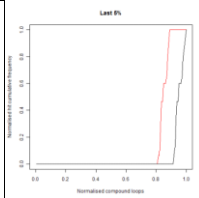
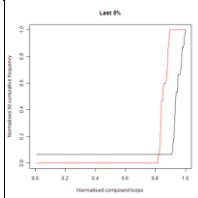
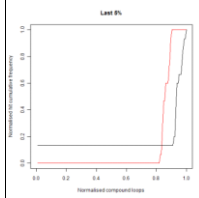
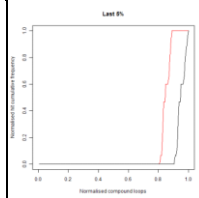
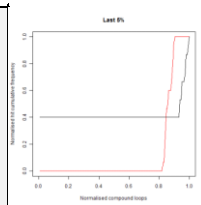
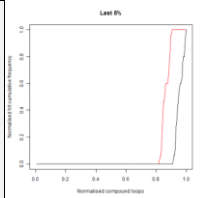
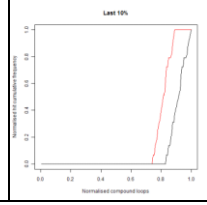
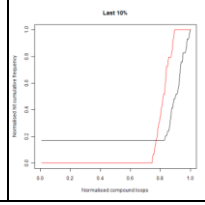
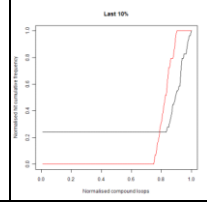
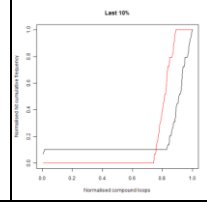
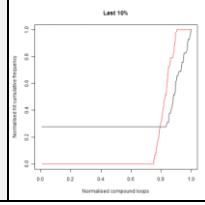
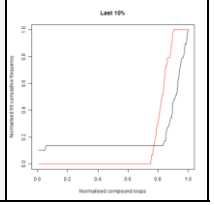
TS6 PfHDFR

	Hs internal seed	External seed parasite/protein						
	HsDHFR	PfDHFR	PvDHFR	TbDHFR	TcDHFR	PvNMT	TbNMT	TcNMT
Single seed, all hits								
Multi seed (Hs, toxic, parasite), all hits								
Last 5%								
Last 10%								

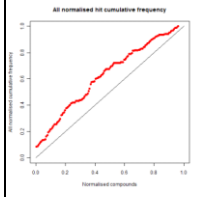
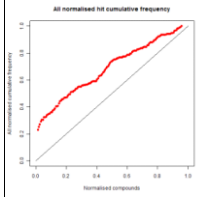
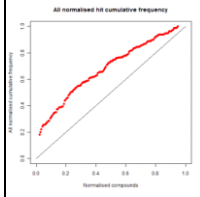
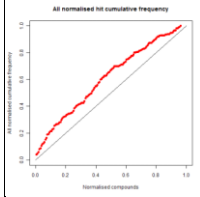
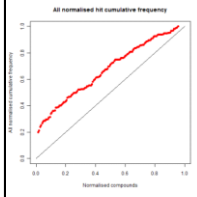
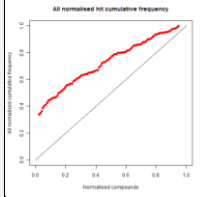
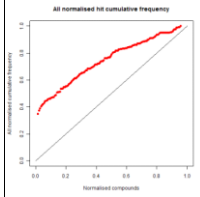
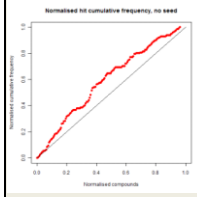
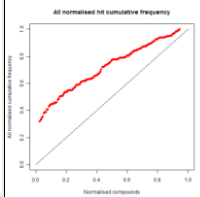
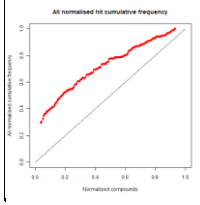
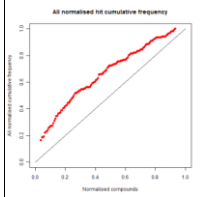
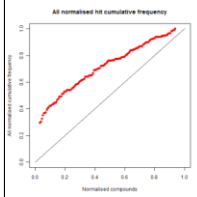
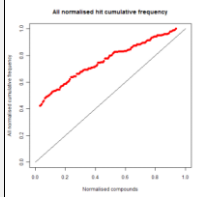
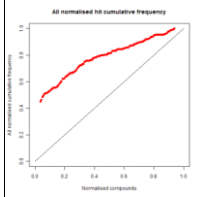
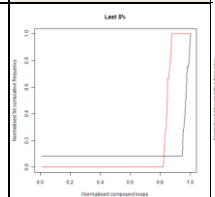
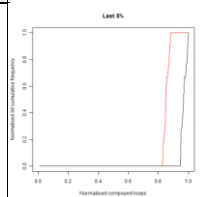
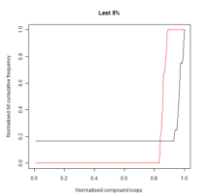
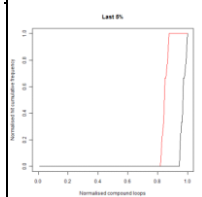
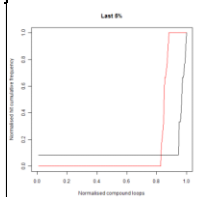
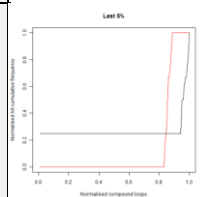
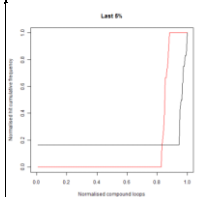
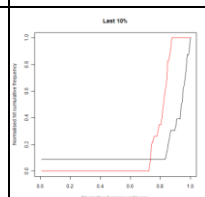
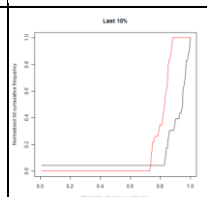
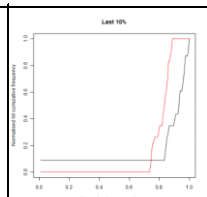
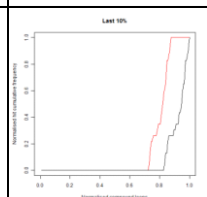
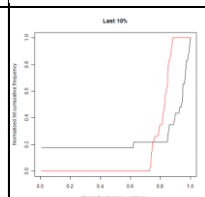
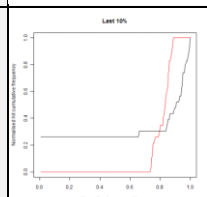
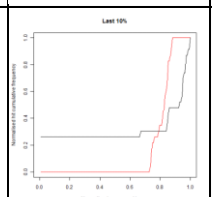
TS7 PvRdhfr

	Hs internal seed	External seed parasite/protein						
	HsDHFR	PfDHFR	PvDHFR	TbDHFR	TcDHFR	PvNMT	TbNMT	TcNMT
Single seed, all hits								
Multi seed (Hs, toxic, parasite), all hits								
Last 5%								
Last 10%								

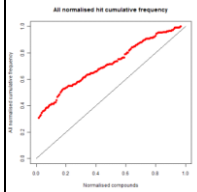
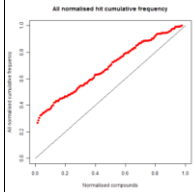
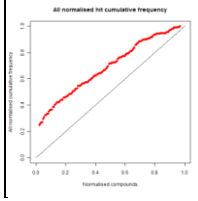
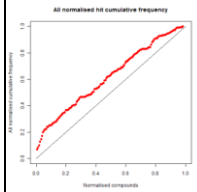
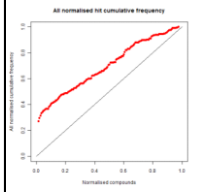
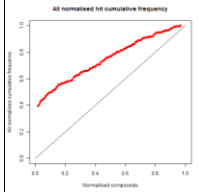
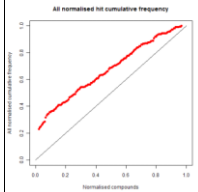
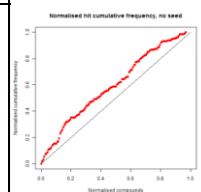
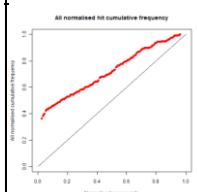
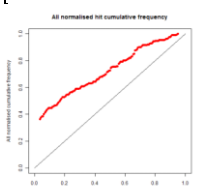
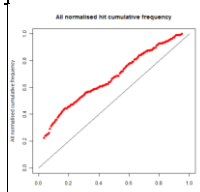
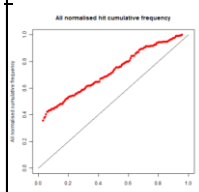
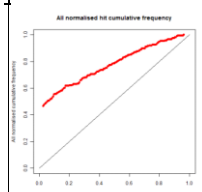
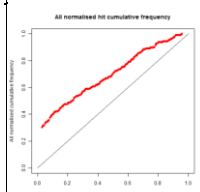
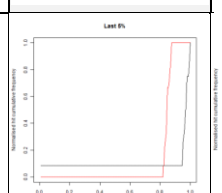
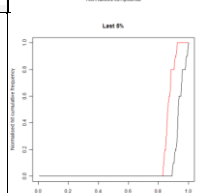
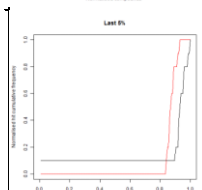
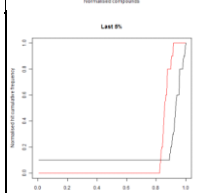
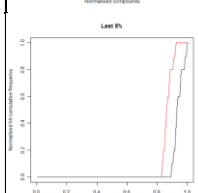
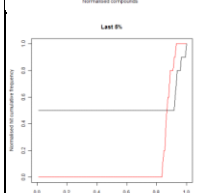
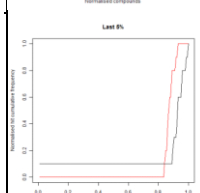
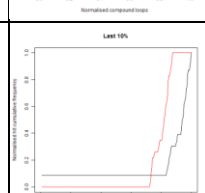
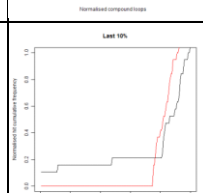
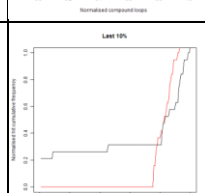
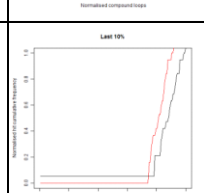
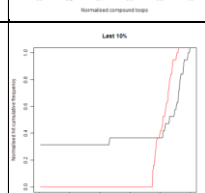
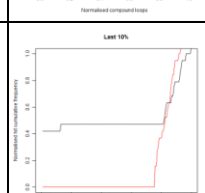
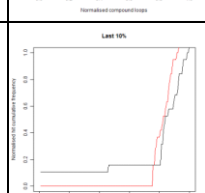
TbNMT

	Hs internal seed	External seed parasite/protein						
	HsDHFR	PfDHFR	PvDHFR	TbDHFR	TcDHFR	PvNMT	TbNMT	TcNMT
Single seed, all hits								
Multi seed (Hs, toxic, parasite), all hits								
Last 5%								
Last 10%								

PvNMT

	Hs internal seed	External seed parasite/protein						
	HsDHFR	PfDHFR	PvDHFR	TbDHFR	TcDHFR	PvNMT	TbNMT	TcNMT
Single seed, all hits								
Multi seed (Hs, toxic, parasite), all hits								
Last 5%								
Last 10%								

TcNMT

	Hs internal seed	External seed parasite/protein						
	HsDHFR	PfDHFR	PvDHFR	TbDHFR	TcDHFR	PvNMT	TbNMT	TcNMT
Single seed, all hits								
Multi seed (Hs, toxic, parasite), all hits								
Last 5%								
Last 10%								

B.5 Deficiency results for combined Transfer Learning/preclustering strategy

Table 1: TL & preclustering (initial TS > 0.6, body TS > 0.4)

Parasite	Protein/ screen	DHFR seeds					NMT seeds		
		Hs	Pf	$Pv_{(TS6)}$	Tb	Tc	Pv	Tb	Tc
Pv	DHFR-TS3		0.73		0.74	0.78	0.76	0.62	0.67
PfR			0.73		0.92		0.77	0.82	
Tb	DHFR-TS4		0.99	0.47		0.75	0.83		0.65
Tc	DHFR-TS5			0.55	0.77		0.63	0.63	0.57
Pv	DHFR-TS6				0.79	0.71	0.64	0.56	0.66
Pf					0.86	0.55	0.60	0.66	0.60
PvR	DHFR-TS7		0.62		0.73		0.63	0.59	0.58
Tb	NMT-1		0.80	0.70	0.81	0.85			0.74
Pv	NMT-1		0.78	0.64	0.71	0.83	0.75		0.51
Tc	NMT-2		0.72	0.66	0.88	0.75	0.54	0.69	
Mean deficiency, all targets			0.77	0.60	0.88	0.75	0.68	0.65	0.62
Mean of deficiencies: 0.70 Mean of core PfDHFR/PvDHFR seed deficiencies: 0.73									

Table 2: Rare category deficiencies (last 5%) for TL & preclustering (initial TS > 0.6, body TS > 0.4)

Parasite	Protein/ screen	DHFR seeds					NMT seeds		
		Hs	Pf	$Pv_{(TS6)}$	Tb	Tc	Pv	Tb	Tc
Pv	DHFR-TS3		0.60		0.64	0.42	0.56	0.40	0.54
PfR			0.51		0.69		0.53	0.60	
Tb	DHFR-TS4		0.60	0.69		0.46	0.59		0.51
Tc	DHFR-TS5			0.29	0.79			0.39	0.48
Pv	DHFR-TS6				0.76	0.41		0.53	0.56
Pf					0.74	0.30		0.41	0.35
PvR	DHFR-TS7		0.46		0.58			0.40	0.48
Tb	NMT-1		0.63		0.74	0.53			0.54
Pv	NMT-1		0.62		0.72	0.62			0.52
Tc	NMT-2		0.65	0.59	0.63	0.65		0.53	
Mean deficiency, all targets			0.58	0.52	0.70	0.48	0.56	0.47	0.50
Mean of deficiencies: 0.54 Mean of core PfDHFR/PvDHFR seed deficiencies: 0.55									

Table 3: Rare category deficiencies (last 10%) for TL & preclustering (initial TS > 0.6, body TS > 0.4)

Parasite	Protein/ screen	DHFR seeds					NMT seeds		
		Hs	Pf	$Pv_{(TS6)}$	Tb	Tc	Pv	Tb	Tc
Pv	DHFR-TS3		0.59		0.61	0.44	0.54	0.48	0.52
PfR			0.50		0.68		0.54	0.58	
Tb	DHFR-TS4		0.55	0.57		0.28	0.40		0.40
Tc	DHFR-TS5			0.42	0.78			0.45	0.38
Pv	DHFR-TS6				0.65	0.48		0.55	0.55
Pf					0.71	0.38		0.42	0.40
PvR	DHFR-TS7		0.49		0.69			0.41	0.57
Tb	NMT-1		0.58		0.71	0.60			0.55
Pv	NMT-1		0.63		0.76	0.58			0.50
Tc	NMT-2		0.55	0.46	0.69	0.44		0.54	
Mean deficiency, all targets			0.56	0.48	0.70	0.46	0.49	0.49	0.48
Mean of deficiencies: 0.52 Mean of core PfDHFR/PvDHFR seed deficiencies: 0.53									

Table 4:TL + preclustering (initial & body TS > 0.4)

Parasite	Protein/ screen	DHFR seeds					NMT seeds		
		Hs	Pf	$Pv_{(TS6)}$	Tb	Tc	Pv	Tb	Tc
Pv	DHFR-TS3		0.63	0.62					
PfR			0.65						
Tb	DHFR-TS4		0.85						
Tc	DHFR-TS5		0.47	0.54					
Pv	DHFR-TS6								
Pf									
PvR	DHFR-TS7		0.70	0.46					
Tb	NMT-1		0.78	0.71					
Pv	NMT-1		0.65	0.71					
Tc	NMT-2		0.70	0.67					
Mean deficiency, all targets			0.68	0.62					
Mean of deficiencies: 0.65 Mean of core PfDHFR/PvDHFR seed deficiencies: 0.69									

Table 5: Rare category deficiencies (last 5%) for TL + preclustering (initial & body TS > 0.4)

Parasite	Protein/ screen	DHFR seeds					NMT seeds		
		Hs	Pf	$Pv_{(TS6)}$	Tb	Tc	Pv	Tb	Tc
Pv	DHFR-TS3		0.68	0.33					
PfR			0.53						
Tb	DHFR-TS4		0.70						
Tc	DHFR-TS5		0.41	0.28					
Pv	DHFR-TS6								
Pf									
PvR	DHFR-TS7		0.50	0.39					
Tb	NMT-1		0.66	0.54					
Pv	NMT-1		0.64	0.49					
Tc	NMT-2		0.66	0.60					
Mean deficiency, all targets			0.60	0.44					
Mean of deficiencies: 0.52 Mean of core PfDHFR/PvDHFR seed deficiencies: 0.58									

Table 6: Rare category deficiencies (last 10%) for TL + preclustering (initial & body TS > 0.4)

Parasite	Protein/ screen	DHFR seeds					NMT seeds		
		Hs	Pf	$Pv_{(TS6)}$	Tb	Tc	Pv	Tb	Tc
Pv	DHFR-TS3		0.62	0.34					
PfR			0.53						
Tb	DHFR-TS4		0.59						
Tc	DHFR-TS5		0.40	0.42					
Pv	DHFR-TS6								
Pf									
PvR	DHFR-TS7		0.49	0.47					
Tb	NMT-1		0.60	0.54					
Pv	NMT-1		0.66	0.58					
Tc	NMT-2		0.55	0.49					
Mean deficiency, all targets			0.56	0.47					
Mean of deficiencies: 0.51 Mean of core PfDHFR/PvDHFR seed deficiencies: 0.55									

Table 7: Subtraction of deficiencies (5.3a – 5.3d)

Parasite	Protein/ screen	DHFR seeds					NMT seeds		
		Hs	Pf	$Pv_{(TS6)}$	Tb	Tc	Pv	Tb	Tc
Pv	DHFR-TS3		-0.14		-0.16	-0.27	-0.16	-0.12	-0.12
PfR			-0.08		-0.04		-0.03	-0.08	
Tb	DHFR-TS4		-0.33	-0.04		-0.30	-0.23		-0.08
Tc	DHFR-TS5			-0.08	-0.11			-0.09	-0.07
Pv	DHFR-TS6				-0.20	-0.16	-0.06	-0.05	-0.07
Pf						-0.12	-0.05	-0.09	-0.06
PvR	DHFR-TS7		-0.13		-0.08		-0.06	-0.07	-0.02
Tb	NMT-1		-0.14	-0.11	-0.09	-0.19			-0.07
Pv	NMT-1		-0.16	-0.01	-0.06	-0.18			-0.01
Tc	NMT-2		-0.10	-0.05	-0.12	-0.15	-0.05	-0.04	
Mean of deficiencies: -0.11 Mean of core PfDHFR/PvDHFR seed deficiencies: -0.13									

Table 8: Subtraction of deficiencies (5.3b – 5.3e)

Parasite	Protein/ screen	DHFR seeds					NMT seeds		
		Hs	Pf	$Pv_{(TS6)}$	Tb	Tc	Pv	Tb	Tc
Pv	DHFR-TS3		0.46		0.43	0.34	0.43	0.28	0.37
PfR			0.39		0.46		0.45	0.44	
Tb	DHFR-TS4		0.63	0.63		0.37	0.64		0.38
Tc	DHFR-TS5			0.23	0.36			0.41	0.35
Pv	DHFR-TS6				0.31	0.34		0.37	0.35
Pf						0.31		0.33	0.27
PvR	DHFR-TS7		0.33		0.31			0.29	0.57
Tb	NMT-1		0.41		0.38	0.35			0.57
Pv	NMT-1		0.52		0.43	0.43			0.43
Tc	NMT-2		0.43	0.38	0.35	0.43		0.44	
Mean of deficiencies: 0.40 Mean of core PfDHFR/PvDHFR seed deficiencies: 0.42									

Table 9: Subtraction of deficiencies (5.3c – 5.3f)

Parasite	Protein/ screen	DHFR seeds					NMT seeds		
		Hs	Pf	$Pv_{(TS6)}$	Tb	Tc	Pv	Tb	Tc
Pv	DHFR-TS3		0.43		0.39	0.40	0.48	0.28	0.40
PfR			0.42		0.43		0.48	0.43	
Tb	DHFR-TS4		0.75	0.71		0.17	0.63		0.62
Tc	DHFR-TS5			0.25	0.37			0.35	0.38
Pv	DHFR-TS6				0.24	0.42		0.45	0.32
Pf						0.32		0.30	0.32
PvR	DHFR-TS7		0.22		0.28			0.20	0.38
Tb	NMT-1		0.34		0.30	0.28			0.41
Pv	NMT-1		0.46		0.39	0.36			0.33
Tc	NMT-2		0.35	0.32	0.34	0.29		0.39	
Mean of deficiencies: 0.38 Mean of core PfDHFR/PvDHFR seed deficiencies: 0.39									

Table 10: Subtraction of deficiencies (5.3a – 5.3g)

Parasite	Protein/ screen	DHFR seeds					NMT seeds		
		Hs	Pf	$Pv_{(TS6)}$	Tb	Tc	Pv	Tb	Tc
Pv	DHFR-TS3		-0.04	-0.33					
PfR			0.00						
Tb	DHFR-TS4		-0.19						
Tc	DHFR-TS5		-0.04	-0.07					
Pv	DHFR-TS6								
Pf									
PvR	DHFR-TS7		-0.18	-0.04					
Tb	NMT-1		-0.12	-0.12					
Pv	NMT-1		-0.03	-0.08					
Tc	NMT-2		-0.08	-0.06					
Mean of deficiencies: -0.11 Mean of core PfDHFR/PvDHFR seed deficiencies: -0.10									

Table 11: Subtraction of deficiencies (5.3b – 5.3h)

Parasite	Protein/ screen	DHFR seeds					NMT seeds		
		Hs	Pf	$Pv_{(TS6)}$	Tb	Tc	Pv	Tb	Tc
Pv	DHFR-TS3		0.38	0.34					
PfR			0.37						
Tb	DHFR-TS4		0.53						
Tc	DHFR-TS5		0.23	0.24					
Pv	DHFR-TS6								
Pf									
PvR	DHFR-TS7		0.29	0.22					
Tb	NMT-1		0.38	0.41					
Pv	NMT-1		0.50	0.45					
Tc	NMT-2		0.42	0.37					
Mean of deficiencies: 0.37 Mean of core PfDHFR/PvDHFR seed deficiencies: 0.39									

Table 12: Subtraction of deficiencies (5.3c – 5.3i)

Parasite	Protein/ screen	DHFR seeds					NMT seeds		
		Hs	Pf	$Pv_{(TS6)}$	Tb	Tc	Pv	Tb	Tc
Pv	DHFR-TS3		0.39	0.22					
PfR			0.39						
Tb	DHFR-TS4		0.71						
Tc	DHFR-TS5		0.18	0.25					
Pv	DHFR-TS6								
Pf									
PvR	DHFR-TS7		0.22	0.22					
Tb	NMT-1		0.32	0.30					
Pv	NMT-1		0.43	0.45					
Tc	NMT-2		0.35	0.29					
Mean of deficiencies: 0.34 Mean of core PfDHFR/PvDHFR seed deficiencies: 0.37									

B.6 Strain-by-strain analysis of Transfer Learning versus endogenous Active Learning algorithms

PvDHFR-TS3

General: The HsDHFR seed deficiency (0.68) is similar to that of *SimplyGreedy* (0.67) and *active k-optimisation* (0.64), indicating there is no significant TL impact from this seed group. In comparison, there is a strong benefit from using the PvDHFR-TS6 seed (0.29), and weaker effects from TcDHFR (0.51) & TbNMT (0.50).

5%/10% rare category: The HsDHFR (0.98/1.01) seed has no significant effect on finding the rare actives. Again, PvDHFR-TS6 is a strong seed (0.67/0.56), and TcDHFR (0.76/0.84) & TbNMT (0.68/0.76) are weaker seeds. In comparison, *active k-optimisation* (0.51/0.47) was a strong identifier of rare category compounds.

PfRdhfr-TS3

General: HsDHFR seed (0.35) versus *SimplyGreedy* (0.88) and *active k-optimisation* (0.91). These results reflect the difficulty in identifying true active compounds against the PfR target (see Chapter 4), and there are no other TL seed groups that can offer obvious benefits.

5%/10% rare category: HsDHFR seed (0.36/0.34), *active k-optimisation* (0.40/0.48). All the other TL seeds (deficiency ≥ 0.89/0.92) fail to predict the rare category compounds. These results further emphasise the lack of true active compounds for this target.

TbDHFR-TS4

General: HsDHFR (0.57), *SimplyGreedy* (0.63), *active k-optimisation* (0.69). PvDHFR (0.43) and TcDHFR (0.45) are strong seeds; TbNMT(0.51) is weaker.

5%/10% rare category: HsDHFR (0.83/1.01), *active k-optimisation* (0.41/0.23). TcDHFR(0.83/0.45) is a strong seed at the 10% level.

TcDHFR-TS5

General: HsDHFR (0.55), *SimplyGreedy* (0.72), *active k-optimisation* (0.73). PfDHFR (0.43) & PvDHFR (0.47) are strong seeds.

5%/10% rare category: HsDHFR (1.02/0.83), *active k-optimisation* (0.44/0.48). PfDHFR(0.64/0.58) and PvDHFR(0.52/0.67) are strong seeds.

PvDHFR-TS6

General: HsDHFR (0.66), *SimplyGreedy* (0.67), *active k-optimisation* (0.62). PfDHFR(0.49) is a strong seed.

5%/10% rare category: HsDHFR (1.06/1.06), *active k-optimisation* (0.45/0.40). PfDHFR (0.75/0.90), TcDHFR (0.75/0.90) and PvNMT (0.83/0.83) are weak seeds.

PfDHFR-TS6

General: HsDHFR (0.54), *SimplyGreedy* (0.74), *active k-optimisation* (0.83). PvDHFR (0.45) and TcDHFR (0.43) are strong seeds.

5%/10% rare category: HsDHFR (0.88/0.86), *active k-optimisation* (0.42/0.44). PvDHFR (0.36/0.57) is a strong seed, and TcDHFR (0.61/0.70) and TcNMT (0.61/0.72) are weak seeds.

PvRdhfr-TS7

General: HsDHFR (0.63), *SimplyGreedy* (0.70), *active k-optimisation* (0.83). PvDHFR (0.42) is a strong seed, and TcDHFR (0.50) is a weaker seed.

5%/10% rare category: HsDHFR (0.95/0.93), *active k-optimisation* (0.57/0.52). PvDHFR (0.61/0.69) and TcDHFR (0.61/0.72) are fairly strong seeds.

TbNMT-1

General: HsDHFR (0.74), *SimplyGreedy* (0.78), *active k-optimisation* (0.72).
PvNMT (0.58) and PvDHFR (0.59) are weak seeds.

5%/10% rare category: HsDHFR (1.12/1.12), *active k-optimisation* (0.43/0.38). PvNMT (0.58/0.80) is a fairly strong seed.

PvNMT-1

General: HsDHFR (0.74), *SimplyGreedy* (0.79), *active k-optimisation* (0.83).
TbNMT (0.53) and TcNMT (0.50) are fairly strong seeds.

5%/10% rare category: HsDHFR (1.05/1.05), *active k-optimisation* (0.42/0.42). TbNMT (0.85/0.83) is a weak seed.

TcNMT-2

General: HsDHFR (0.56), *SimplyGreedy* (0.80), *active k-optimisation* (0.80).
PvNMT (0.49) is a weak seed.

5%/10% rare category: HsDHFR (0.76/0.73), *active k-optimisation* (0.40/0.36). PvNMT (0.54/0.57) is a weak seed.

Appendix C

C.1 Mass & cherrypick screen .csv files, column contents

No.	Data name	Contents
A/1	id_plate_instance	numeric, plate number
B/2	well_row	letter, always A-P
C/3	well_col	numeric, always 1-24
D/4	id_mixture	string, two entries for each filled well, one entry for each negative control, no entries for empty wells, uses library name/code for substances on test
E/5	id_chemical	numeric, Eve ID for chemicals in well, two entries for compounds (compound+DMSO), one entry for negative controls (#1313)
F/6	name	string, full compound name (or DMSO)
G/7	smiles	SMILES code for F/6
H/8	cherry_testname	alphanumeric
I/9	cherry_hadError	numeric, should be zero
J/10	cherry_startvalue	numeric, initial fluorescence
K/11	cherry_endvalue	numeric, final fluorescence
L/12	cherry_miylagtime	numeric
M/13	cherry_doubletime	numeric
N/14	sapphire_testname	alphanumeric
O/15	sapphire_hadError	numeric, should be zero
P/16	sapphire_startvalue	numeric, initial fluorescence
Q/17	sapphire_endvalue	numeric, final fluorescence
R/18	sapphire_miylagtime	numeric
S/19	sapphire_doubletime	numeric
T/20	venus_testname	alphanumeric
U/21	venus_hadError	numeric, should be zero
V/22	venus_startvalue	numeric, initial fluorescence
W/23	venus_endvalue	numeric, final fluorescence
X/24	venus_miylagtime	numeric
Y/25	venus_doubletime	numeric
Z/26	id_plate_layout	numeric
AA/27	id_study	numeric
AB/28	vol_cmpd_nl	numeric, for confirmation screen use
AC/29	cmpd_conc_um	numeric, for confirmation screen use

C.2 Hardware used during data analysis and simulations

Early prototype AL simulations, developed based on Kurt De Grave's *active k-optimisation* algorithm (coded in Octave as multiSelOpt.m), were conducted using:

(i) "Cledwall-testing" (processor 1: AMD Athlon MP 2000+ 1.67 GHz; processor 2: AMD Athlon MP 1000 – 1733 GHz; 2GB RAM).
(ii) "Cledwall" (Intel Xeon quad core 2.80 GHz; 4 GB RAM).
(iii) A desktop PC (Intel Core 2 Duo E8400 3.00 GHz; 4 GB RAM) running the Cygwin simulator.

All three of these hardware setups ran into problems when attempting multiple loop simulations: Cledwall-testing would run three loops then slow rapidly due to insufficient memory; similar but less severe problems were seen with the desktop PC and Cledwall. It was also discovered that the 64-bit version of Octave is needed to process the larger data sets generated as the simulation progresses; this restricted the PC Cygwin Linux simulator as it did not support this software (August 2011).

The Beowulf cluster (up to 40 units, Intel Pentium4 2.8 GHz with 1GB RAM) was also tentatively proposed as a possible resource for running simulations. In addition to likely problems with computing capacity, this cluster didn't have appropriate software installed or any maintenance service contract.

After tentatively exploring a couple of other external resources (cloud computing & HPC Wales), access to a computing facility (PINAC) at the Katholieke Universiteit Leuven was finally provided through our collaborator, Kurt de Grave.

The PINAC cluster has developed over the period in use, with five older machines being replaced by 10 newer models shortly before the end of the set of simulations:

No.	Specification	Oct 2011 to Feb 2012	Feb 2012 onwards
5	Dual core E6600 processors (2.4 GHz, 4 MB cache), 4 GB RAM, 160 GB hard drive	×	
10	Quad core Q9550 processors (2.83 GHz, 12 MB cache), 8 GB RAM, 320 GB hard drive	×	×
10	Intel Core i7-2600 3.4 GHz, 16 GB RAM		×

The PINAC facility was used for all work on the *active k-optimisation* strategy.

All other AL simulations were conducted using the HPC cluster run by IBERS, Aberystwyth University. This facility has 152 processor cores available, with three discrete configurations:

3 × AMD nodes

AMD Opteron Processor 6220 @ 3 GHz

32 cores

100 GB memory available per node

3 × Intel nodes

Intel(R) Xeon(R) CPU X5647 @2.93 GHz

8 cores

190GB memory available per node

1 × large memory node

AMD Opteron Processor 6220 @ 3 GHz

32 cores

500 GB memory

C.3 Software used during data analysis and simulations

The data analysis steps were conducted using:

- Microsoft Office Excel 2007.
- WEKA (version 3.6.2) J48 decision trees.
- R (version 2.12).

The *active k-optimisation* simulations used:

- R (version 2.12).
- OpenBabel (version 2.2.3). Note: some compatibility problems were encountered with later versions when used in conjunction with the Octave routine provided by Leuven.
- Java.
- Octave (64 bit Linux version 3.2.3 “x86_64-pc-linux-gnu” in Leuven; 32 bit Windows version 3.2.4 for testing purposes on a PC, in combination with the Cygwin Linux simulator).

All other AL simulations used:

- R 2.13.1 (including the following libraries: fingerprint 3.4.7, rcdk 3.1.7, rcdklibs 1.4.7, rJava 0.9-3, rpubchem 1.4.3, car 2.0-11, rCurl 1.6-10.1, XML 3.4-2.2, bitops 1.0-4.1, methods 2.13.1, png 0.1-4, iterators 1.0.5, xtable 1.7-0, RUnit 0.4.26)

References

E. Alpaydin, *Introduction to machine learning*, The MIT Press, 2004.

Christopher Arico-Muendel, Paolo A Centrella, Brooke D Contonio, Barry A Morgan, Gary ODonovan, Christopher L Paradise, Steven R Skinner, Barbara Sluboski, Jennifer L Svendsen, Kerry F White, et al., *Antiparasitic activities of novel, orally available fumagillin analogs*, Bioorganic & medicinal chemistry letters **19** (2009), no. 17, 5128-5131.

Tomasz Arodz and Arkadiusz Z Dudek, *Multivariate modeling and analysis in drug discovery*, Current Computer-Aided Drug Design **3** (2007), no. 4, 240-247.

Alyson M Auli, John H Adams, Michael T O'Neil, and Qin Cheng, *Defining the role of mutations in plasmodium vivax dihydrofolate reductase-thymidylate synthase gene using an episomal plasmodium falciparum transfection system*, Antimicrobial agents and chemotherapy **54** (2010), no. 9, 3927-3932.

O. Babel, *The Open Source Chemistry Toolbox*.

Nicola Baker, Harry P de Koning, Pascal Mäser, and David Horn, *Drug resistance in african trypanosomiasis: the melarsoprol and pentamidine story*, Trends in parasitology (2013).

Yoram Baram, Ran El-Yaniv, and Kobi Luz, *Online choice of active learning algorithms*, The Journal of Machine Learning Research **5** (2004), 255-291.

SJ Barrett and WB Langdon, *Advances in the application of machine learning techniques in drug discovery, design and development*, Applications of Soft Computing, Springer, 2006, pp. 99-110.

Elizabeth Bilsland, Pnar Pir, Alex Gutteridge, Alexander Johns, Ross D King, and Stephen G Oliver, *Functional expression of parasite drug targets and their human orthologs in yeast*, PLoS neglected tropical diseases **5** (2011), no. 10, e1320.

Elizabeth Bilsland, Andrew Sparkes, Kevin Williams, Harry J Moss, Michaela de Clare, Pinar Pir, Jem Rowland,Wayne Aubrey, Ron Pateman, Mike Young, et al., *Yeast-based automated high-throughput screens to identify anti-parasitic lead compounds*, Open biology **3** (2013), no. 2.

C.M. Bishop et al., *Pattern recognition and machine learning*, Springer New York, 2006.

Mark S Boguski, Kenneth D Mandl, and Vikas P Sukhatme, *Repurposing with a difference*, Science **324** (2009), no. 5933, 1394.

B. Bringmann and A. Karwath, *Frequent smiles*, Lernen, Wissensentdeckung und Adaptivitat, Workshop GI Fachgruppe Maschinelles Lernen, LWA, Citeseer, 2004.

Robert Burbidge, Matthew Trotter, B Buxton, and S Holden, *Drug design by machine learning: support vector machines for pharmaceutical data analysis*, Computers & Chemistry **26** (2001), no. 1, 5-14.

Mark S Butler and Antony D Buss, *Natural products the future scaffolds for novel antibiotics?*, Biochemical pharmacology **71** (2006), no. 7, 919-929.

Rich Caruana, *Multitask learning*, Machine learning **28** (1997), no. 1, 41-75.

Jaya Chakravarty and Shyam Sundar, *Drug resistance in leishmaniasis*, Journal of global infectious diseases **2** (2010), no. 2, 167.

Curtis R Chong and David J Sullivan, *New uses for old drugs*, Nature **448** (2007), 645-646.

Thomas Cover and Peter Hart, *Nearest neighbor pattern classification*, Information Theory, IEEE Transactions on **13** (1967), no. 1, 21-27.

Alan F Cowman, Mary J Morry, Beverly A Biggs, GA Cross, and Simon J Foote, *Amino acid changes linked to pyrimethamine resistance in the dihydrofolate reductase-thymidylate synthase gene of plasmodium falciparum*, Proceedings of the National Academy of Sciences **85** (1988), no. 23, 9109-9113.

Dennis D Cox and Susan John, *Sdo: A statistical method for global optimization*, Multidisciplinary design optimization: state of the art (1997), 315-329.

Sanjoy Dasgupta and Daniel Hsu, *Hierarchical sampling for active learning*, Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 208-215.

Jesse Davis and Mark Goadrich, *The relationship between precision-recall and roc curves*, Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 233-240.

Kurt De Grave, Jan Ramon, and Luc De Raedt, *Active learning for high throughput screening*, Discovery Science, Springer, 2008, pp. 185-196.

Wanderley de Souza, Tecia Maria Ulisses de Carvalho, and Emile Santos Barrias, *Review on trypanosoma cruzi: host cell interaction*, International journal of cell biology **2010** (2010).

Arjen M Dondorp, Francois Nosten, Poravuth Yi, Debashish Das, Aung Phae Phyo, Joel Tarning, Khin Maung Lwin, Frederic Ariey, Warunee Hanpithakpong, Sue J Lee, et al., *Artemisinin resistance in plasmodium falciparum malaria*, New England Journal of Medicine **361** (2009), no. 5, 455-467.

J. Drews, *Drug discovery: a historical perspective*, Science **287** (2000), no. 5460, 1960.

Arkadiusz Z Dudek, Tomasz Arodz, and Jorge Galvez, *Computational methods in developing quantitative structure-activity relationships (qsar): a review*, Combinatorial chemistry & high throughput screening **9** (2006), no. 3, 213-228.

Nick Feasey, Mark Wansbrough-Jones, David CW Mabey, and Anthony W Solomon, *Neglected tropical diseases*, British medical bulletin **93** (2010), no. 1, 179-200.

Michael A Fischbach and Christopher T Walsh, *Antibiotics for emerging pathogens*, Science **325** (2009), no. 5944, 1089-1093.

Julie A Frearson, Stephen Brand, Stuart P McElroy, Laura AT Cleghorn, Ondrej Smid, Laste Stojanovski, Helen P Price, M Lucia S Guther, Leah S Torrie, David A Robinson, et al., *N-myristoyltransferase inhibitors as new leads to treat sleeping sickness*, Nature **464** (2010), no. 7289, 728-732.

Zoubin Ghahramani, *Unsupervised learning*, Advanced Lectures on Machine Learning, Springer, 2004, pp. 72-112.

Jahan B Ghasemi and Fereshteh Shiri, *Molecular docking and 3d-qsar studies of falcipain inhibitors using comfa, comsia, and open3dqsar*, Medicinal Chemistry Research **21** (2012), no. 10, 2788-2806.

P. J. Goodford, *Drug design by the method of receptor fit*, Journal of Medicinal Chemistry **27** (1984), no. 5, 557-564, 59 AMER CHEMICAL SOC WASHINGTON SP667.

R. Guha, M.T. Howard, G.R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, and E.L. Willighagen, *The Blue Obelisk-interoperability in chemical informatics*, J. Chem. Inf. Model **46** (2006), no. 3, 991-998.

R. Guha, *Chemical informatics functionality in r*, Journal of Statistical Software **18** (2007), no. 5, 1-16.

R. Guha, *Package rcdk*, (2013).

Felix Hammann and Juergen Drewe, *Decision tree models for data mining in hit discovery*, Expert Opinion on Drug Discovery **7** (2012), no. 4, 341-352.

Shuli Han, Bo Yuan, and Wenhuang Liu, *Rare class mining: progress and prospect*, Pattern Recognition, 2009. CCPR 2009. Chinese Conference on, IEEE, 2009, pp. 1-5.

John A Hartigan and Manchek A Wong, *Algorithm as 136: A k-means clustering algorithm*, Applied statistics (1979), 100-108.

Alan L Harvey, Rachel L Clark, Simon P Mackay, and Blair F Johnston, *Current strategies for drug discovery through natural products*, Expert opinion on drug discovery **5** (2010), no. 6, 559-568.

Jingrui He and Jaime G Carbonell, *Nearest-neighbor-based active learning for rare category detection*, Advances in Neural Information Processing Systems, 2007, pp. 633-640.

Stephen R Heller and Alan D McNaught, *The iupac international chemical identifier (inchi)*, Chemistry International **31** (2009), no. 1, 7.

Ruili Huang, Noel Southall, Yuhong Wang, Adam Yasgar, Paul Shinn, Ajit Jadhav, Dac-Trung Nguyen, and Christopher P Austin, *The ncgc pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics*, Science translational medicine **3** (2011), no. 80, 80ps16.

Anil K Jain, M Narasimha Murty, and Patrick J Flynn, *Data clustering: a review*, ACM computing surveys (CSUR) **31** (1999), no. 3, 264-323.

CA James, DWeininger, and J Delaney, *Fingerprints screening and similarity. Daylight theory manual*, Daylight chemical information systems inc., 1997.

Catherine E James, Amanda L Hudson, and Mary W Davey, *Drug resistance mechanisms in helminths: is it survival of the fittest?*, Trends in parasitology **25** (2009), no. 7, 328-335.

Craig A James, D Weininger, and J Delany, *Daylight theory manual*, Daylight chemical information systems **3951** (1995).

Donald R Jones, Matthias Schonlau, and William J Welch, *Efficient global optimization of expensive black-box functions*, Journal of Global optimization **13** (1998), no. 4, 455-492.

A. Karwath and L. De Raedt, *Smirep: Predicting chemical activity from smiles*, J. Chem. Inf. Model **46** (2006), no. 6, 2432-2444.

M.J. Keiser, V. Setola, J.J. Irwin, C. Laggner, A.I. Abbas, S.J. Hufeisen, N.H. Jensen, M.B. Kuijer, R.C. Matos, T.B. Tran, et al., *Predicting new molecular targets for known drugs*, Nature **462** (2009), no. 7270, 175-181.

György M Keserü and Gergely M Makara, *Hit discovery and hit-to-lead approaches*, Drug discovery today **11** (2006), no. 15, 741-748.

R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, A. Sparkes, K. E. Whelan, and A. Clare, *The automation of science*, Science **324** (2009), no. 5923, 85-89.

R. D. King, K. E. Whelan, F. M. Jones, P. G. K. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, and S. G. Oliver, *Functional genomic hypothesis generation and experimentation by a robot scientist*, Nature **427** (2004), no. 6971, 247-252.

Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, et al., *Drugbank 3.0: a comprehensive resource for omics research on drugs*, Nucleic acids research **39** (2011), no. suppl 1, D1035-D1041.

Frank E Koehn and Guy T Carter, *The evolving role of natural products in drug discovery*, Nature Reviews Drug Discovery **4** (2005), no. 3, 206-220.

Palangpon Kongsaeree, Puttapol Khongsuk, Ubolsree Leartsakulpanich, Penchit Chitnumsub, Bongkoch Tarnchompoo, Malcolm D Walkinshaw, and Yongyuth Yuthavong, *Crystal structure of dihydrofolate reductase from plasmodium vivax: pyrimethamine displacement linked with mutation-induced resistance*, Proceedings of the National Academy of Sciences of the United States of America **102** (2005), no. 37, 13046-13051.

Hongdong Li, Yizeng Liang, and Qingsong Xu, *Support vector machines and its applications in chemistry*, Chemometrics and Intelligent Laboratory Systems **95** (2009), no. 2, 188-198.

C. A. Lipinski, *Drug-like properties and the causes of poor solubility and poor permeability*, Journal of Pharmacological and Toxicological Methods **44** (2000), no. 1, 235-249.

C. A. Lipinski, *The anti-intellectual effects of intellectual property*, Current Opinion in Chemical Biology **10** (2006), no. 4, 380-383.

C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*, Advanced Drug Delivery Reviews **23** (1997), no. 1-3, 3-25.

Felipe Lira, Pedro S Perez, Jose A Baranauskas, and Sergio R Nozawa, *Prediction of antimicrobial activity of synthetic peptides by a decision tree model*, Applied and Environmental Microbiology **79** (2013), no. 10, 3156-3159.

K. McConway, MC Jones, and PC Taylor, *Statistical modelling using GENSTAT*, Arnold, London, 1999.

Georgia B McGaughey, Robert P Sheridan, Christopher I Bayly, J Chris Culberson, Constantine Kreatsoulas, Stacey Lindsley, Vladimir Maiorov, Jean Francois Truchon, and Wendy D Cornell, *Comparison of topological, shape, and docking methods in virtual screening*, Journal of chemical information and modeling **47** (2007), no. 4, 1504-1519.

Malcolm J McGregor and Peter V Pallai, *Clustering of large databases of compounds: Using the mdl keys as structural descriptors*, Journal of chemical information and computer sciences **37** (1997), no. 3, 443-448.

Laura M McMurry, Margret Oethinger, and Stuart B Levy, *Triclosan targets lipid synthesis*, Nature **394** (1998), no. 6693, 531-532.

Peter C Melby, Richard D Kreutzer, Diane McMahon-Pratt, Albert A Gam, and Franklin A Neva, *Cutaneous leishmaniasis: review of 59 cases seen at the national institutes of health*, Clinical infectious diseases **15** (1992), no. 6, 924-937.

Sandra D Melman, Michelle L Steinauer, Charles Cunningham, Laura S Kubatko, Ibrahim N Mwangi, Nirvana Barker Wynn, Martin W Mutuku, Diana MS Karanja, Daniel G Colley, Carla L Black, et al., *Reduced susceptibility to praziquantel among naturally occurring kenyan isolates of schistosoma mansoni*, PLoS neglected tropical diseases **3** (2009), no. 8, e504.

J.L. Melville, J.F. Riley, and J.D. Hirst, *Similarity by compression*, J. Chem. Inf. Model **47** (2007), no. 1, 25-33.

Paul AM Michels, Frederic Bringaud, Murielle Herman, and Veronique Hannaert, *Metabolic functions of glycosomes in trypanosomatids*, Biochimica et Biophysica Acta (BBA)-Molecular Cell Research **1763** (2006), no. 12, 1463-1477.

Mary Moran, *Global funding of new products for neglected tropical diseases*, The Causes and Impacts of Neglected Tropical and Zoonotic Diseases: Opportunities for Integrated Intervention Strategies: Workshop Summary, National Academies Press, 2011, p. 388.

Steve Morgan, Paul Grootendorst, Joel Lexchin, Colleen Cunningham, and Devon Greyson, *The cost of drug development: a systematic review*, Health Policy **100** (2011), no. 1, 4-17.

Alexis Nzila, Matthias Rottmann, Penchit Chitnumsub, Stevens M Kiara, Sumalee Kamchonwongpaisan, Cherdsak Maneeruttanarungroj, Supannee Taweechai, Bryan KS Yeung, Anne Goh, Suresh B Lakshminarayana, et al., *Preclinical evaluation of the antifolate qn254 as an antimalarial drug candidate*, Antimicrobial agents and chemotherapy **54** (2010), no. 6, 2603-2610.

Robert H O'Neil, Ryan H Lilien, Bruce R Donald, Robert M Stroud, and Amy C Anderson, *Phylogenetic classiffcation of protozoa based on the structure of the linker domain in the bifunctional enzyme, dihydrofolate reductase-thymidylate synthase*, Journal of Biological Chemistry **278** (2003), no. 52, 52980-52987.

Sinno Jialin Pan and Qiang Yang, *A survey on transfer learning*, Knowledge and Data Engineering, IEEE Transactions on **22** (2010), no. 10, 1345-1359.

Dinesh V Patel and Eric M Gordon, *Applications of small-molecule combinatorial chemistry to drug discovery*, Drug Discovery Today **1** (1996), no. 4, 134-144.

Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht, *How to improve r&d productivity: the pharmaceutical industry's grand challenge*, Nature reviews Drug discovery **9** (2010), no. 3, 203-214.

Dan Pelleg and Andrew W Moore, *Active learning for anomaly and rare category detection*, Advances in Neural Information Processing Systems, 2004, pp. 1073-1080.

David S Peterson, David Walliker, and Thomas E Wellems, *Evidence that a point mutation in dihydrofolate reductase-thymidylate synthase confers resistance to pyrimethamine in falciparum malaria*, Proceedings of the National Academy of Sciences **85** (1988), no. 23, 9114-9118.

Francisco J Prado-Prado, Xerardo García-Mera, and Humberto González-Díaz, *Multi-target spectral moment qsar versus ann for antiparasitic drugs against diff*

erent parasite species, Bioorganic & medicinal chemistry **18** (2010), no. 6, 2225-2231.

Robert G Ridley, *Chemotherapeutic hope on the horizon for plasmodium vivax malaria?*, Proceedings of the National Academy of Sciences **99** (2002), no. 21, 13362-13364.

Burr Settles, *Active learning literature survey*, University of Wisconsin, Madison (2010).

Worachart Sirawaraporn, Tanajit Sathitkul, Rachada Sirawaraporn, Yongyuth Yuthavong, and Daniel V Santi, *Antifolate-resistant mutants of plasmodium falciparum dihydrofolate reductase*, Proceedings of the National Academy of Sciences **94** (1997), no. 4, 1124-1129.

Andrew Sparkes, Wayne Aubrey, Emma Byrne, Amanda Clare, Muhammed N Khan, Maria Liakata, Magdalena Markham, Jem Rowland, Larisa N Soldatova, Kenneth E Whelan, et al., *Review towards robot scientists for autonomous scientifc discovery*, Autom Exp **2** (2010).

Ola Spjuth, Jonathan Alvarsson, Arvid Berg, Martin Eklund, Stefan Kuhn, Carl M□asak, Gilleain Torrance, Johannes Wagener, Egon Willighagen, Christoph Steinbeck, et al., *Bioclipse 2: A scriptable integration platform for the life sciences*, BMC bioinformatics **10** (2009), no. 1, 397.

Ola Spjuth, Valentin Georgiev, Lars Carlsson, Jonathan Alvarsson, Arvid Berg, Egon Willighagen, Jarl ES Wikberg, and Martin Eklund, *Bioclipse-r: integrating management and visualization of life science data with statistical analysis*, Bioinformatics **29** (2013), no. 2, 286-289.

Ola Spjuth, Tobias Helmus, Egon L Willighagen, Stefan Kuhn, Martin Eklund, Johannes Wagener, Peter Murray-Rust, Christoph Steinbeck, Jarl ES Wikberg, et al., *Bioclipse: an open source workbench for chemo-and bioinformatics*, BMC bioinformatics **8** (2007), no. 1, 59.

Dietmar Steverding, *The development of drugs for treatment of sleeping sickness: a historical review*, Parasit Vectors **3** (2010), no. 1, 15.

Claudia Suenderhauf, Felix Hammann, and Jörg Huwyler, *Computational prediction of blood-brain barrier permeability using decision tree induction*, Molecules **17** (2012), no. 9, 10429-10445.

Namita Surolia and Avadhesha Surolia, *Triclosan offers protection against blood stages of malaria by inhibiting enoyl-acp reductase of plasmodium falciparum*, Nature medicine **7** (2001), no. 2, 167-173.

TT Tanimoto, *IBM Internal Report 1957*, 1957.

I.V. Tetko, *Computing chemistry on the web*, Drug discovery today **10** (2005), no. 22, 1497-1500.

I.V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V.A. Palyulin, E.V. Radchenko, N.S. Zerov, A.S. Makarenko, et al., *Virtual computational chemistry laboratory-design and description*, Journal of computer-aided molecular design **19** (2005), no. 6, 453-463.

Michel Tibayrenc, Finn Kjellberg, and Francisco J Ayala, *A clonal theory of parasitic protozoa: the population structures of entamoeba, giardia, leishmania, naegleria, plasmodium, trichomonas, and trypanosoma and their medical and taxonomical consequences*, Proceedings of the National Academy of Sciences **87** (1990), no. 7, 2414-2418.

A. A. Toropov and E. Benfenati, *Qsar modelling of aldehyde toxicity against a protozoan, tetrahymena pyriformis by optimization of correlation weights of nearest neighboring codes*, Journal of Molecular Structure-Theochem **679** (2004), no. 3, 225-228.

A. A. Toropov, E. Benfenati, *Optimisation of correlation weights of smiles invariants for modelling oral quail toxicity*, European Journal of Medicinal Chemistry **42** (2007), no. 5, 606-613.

A. A. Toropov and T. W. Schultz, *Prediction of aquatic toxicity: Use of optimization of correlation weights of local graph invariants*, Journal of Chemical Information and Computer Sciences **43** (2003), no. 2, 560-567, 6th International Conference on Chemical Structures JUN, 2002 NOORDWIJKERHOUT, NETHERLANDS.

A. A. Toropov, A. P. Toropova, and E. Benfenati, *Simplifed molecular input line entry system-based optimal descriptors: Quantitative structure-activity relationship modeling mutagenicity of nitrated polycyclic aromatic hydrocarbons*, Chemical Biology & Drug Design **73** (2009), no. 5, 515-525.

Paolo Tosco and Thomas Balle*, Open3dqsar: a new open-source software aimed at high-throughput chemometric analysis of molecular interaction fields*, Journal of molecular modeling **17** (2011), no. 1, 201-208.

Alexander Tropsha, *Best practices for qsar model development, validation, and exploitation*, Molecular Informatics **29** (2010), no. 6-7, 476-488.

L. G. Valiant, *A theory of the learnable*, Communications of the Acm **27** (1984), no. 11, 1134-1142.

Pieter Vandezande, Lieven EM Gevers, Johan S Paul, Ivo FJ Vankelecom, and Pierre A Jacobs, *High throughput screening for rapid development of membranes and membrane processes*, Journal of membrane science **250** (2005), no. 1, 305-310.

M. K. Warmuth, J. Liao, G. Ratsch, M. Mathieson, S. Putta, and C. Lemmen, *Active learning with support vector machines in the drug discovery process*, Journal of Chemical Information and Computer Sciences **43** (2003), no. 2, 667-673, 6th International Conference on Chemical Structures JUN, 2002 NOORDWIJKERHOUT, NETHERLANDS.

Manfred K Warmuth, Gunnar Ratsch, Michael Mathieson, Jun Liao, and Christian Lemmen, *Active learning in the drug discovery process*, NIPS, 2001, pp. 1449-1456.

D.Weininger, *SMILES 1. Introduction and encoding rules*, J. Chem. Inf. Comput. Sci **28** (1988), 31.

D. Weininger, A. Weininger, and J.L. Weininger, *SMILES. 2. Algorithm for generation of unique SMILES notatio*n, Journal of Chemical Information and Computer Sciences **29** (1989), no. 2, 97-101.

Peter Willett, John M Barnard, and Geoffrey M Downs, *Chemical similarity searching*, Journal of Chemical Information and Computer Sciences **38** (1998), no. 6, 983-996.

David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey, *Drugbank: a comprehensive resource for in silico drug discovery and exploration*, Nucleic acids research **34** (2006), no. suppl 1, D668-D672.

I.H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann Pub, 2005.

Min Yu, TR Kumar, Louis J Nkrumah, Alida Coppi, Silke Retzla, Celeste D Li, Brendan J Kelly, Pedro A Moura, Viswanathan Lakshmanan, Joel S Freundlich, et al., *The fatty acid biosynthesis enzyme fabi plays a key role in the development of liver-stage malarial parasites*, Cell host & microbe **4** (2008), no. 6, 567-578.

Jirundon Yuvaniyama, Penchit Chitnumsub, Sumalee Kamchonwongpaisan, Jarunee Vanichtanankul, Worachart Sirawaraporn, Paul Taylor, Malcolm D Walkinshaw, and Yongyuth Yuthavong, *Insights into antifolate resistance from malarial dhfr-ts structures*, Nature Structural & Molecular Biology **10** (2003), no. 5, 357-365.